SESSION

ONTOLOGIES

Chair(s)

TBA

Semantic Design Patterns Using the OWL Domain Profile

Rishi Kanth Saripalle, and Steven A Demurjian
Department of Computer Science & Engineering, University of Connecticut, Storrs, CT, USA

Abstract— *Design patterns in software engineering are templates garnered from generalizing proven solutions that can then be applied to other applications with similar needs; patterns allows developers to quickly structure their software based on a combination of its information and behavioral characteristics. However, the research, design, development, and usage of design patterns in ontologies is at a novice level. Patterns as applied to ontology design and development have the potential to promote reuse across various domain ontologies. Hence, the creation of reusable semantic design patterns and their inclusion into ontologies should enhance both the semantics and the reusability of ontologies. This paper proposes an approach to represent semantic design patterns using the OWL Web Ontology Language Domain Profile feature by integrating patterns into the domain ontology.*

Keywords - Knowledge Patterns, ODP, OWL Metamodel

1 Introduction

A *design pattern* in software engineering can be defined as "*a description or template as how to solve a reoccurring problem that can be used in many different situations.*"[1]. Essentially, design patterns are templates, but not furnished domain models, which can be used for capturing domain data or transformed into executable code. As a result, design patterns are an architectural concept (e.g., the structure and behavior of a system or component) which are realized in programming languages such as Java, C++, etc. The ability of patterns to facilitate *modularity, flexibility,* and *ease of transformation* across heterogeneous systems, have gained them wide acceptance in academic research and industry.

Design patterns typically show *relationships* and *interactions* between classes or objects at a conceptual level, for domain independent structure and behavior. A pattern provides a body of knowledge on a particular structure thereby communicating insight on a portion or component of a solution; the idea is that we can leverage patterns that are generalized from prior solutions and experience to build solutions more effectively. While patterns are dominant in the software engineering community, they have also influenced other research areas like computational algorithm design and application execution behavior [12]; patterns have much to offer beyond their original usage if one generalizes them even more to apply to other research areas.

Once such research area is knowledge representation, where there has been work that begins to apply patterns [2, 8, 9]. Knowledge representation primarily focuses on the development of *ontologies* which captures the semantic agreement a conceptualization of a domain. The question is whether the domain knowledge encapsulated by the ontology can be generalized (patterned) to be reused in multiple settings. For example, in the medical domain, each system (e.g., electronic medical record, prescription system, personal health record, etc.) may all have their own ontologies to manage diseases, symptoms, etc. The results are often incompatibility among ontologies for the same "data" that is not easily reusable. This is often due to an ontology designer not having a tool such design patterns to assist in structuring these larger and complex ontologies. Semantic design patterns which were previously overlooked while designing and developing ontologies has gained importance in the recent times, as they provide generalized structure, which explicitly modularizes conceptualization from domain concepts and rich axioms giving meaning to the structure [2]. For example, rather than encoding a theory about different *electric circuits* which can exist/co-exist in an electronic appliance, we can initially design *directed graphs* and later develop a procedure as how to impose or correlate the circuit to the graphs[2]. In a medical domain, we can leverage patterns to develop ontologies for diseases, symptoms, etc., that are more easily reused to facilitate sharing of medical information In both examples, patterns provide modular structures which are reusable; encourage knowledge modeling rather than encoding; and eases integration issues since rules built at the pattern level have to be followed by lower levels. Further, ontology can be viewed as collection of patterns which are connected with each other instead of behemoth logical structure.

In this paper, we propose the development of semantic design patterns by identifying domain specific abstract concepts and leveraging the OWL Domain Profile (ODP) to develop a pattern. The paper is organized as follows: Section 2 provides background information on design patterns, and OWL Domain Profile, and a review of prior and ongoing efforts on utilizing patterns for ontologies; Section 3 proposes and demonstrates a technique of utilizing design patterns with ODP; and Section 4 concludes the paper and discusses ongoing work.

2 Background and Motivation

Design patterns [1] captures the structure and semantics of a component level solution that has been generalized from well-proven solutions taken from different domains; as a result, such a pattern can be adopted to work in different applications that have the similar requirements of the pattern. The *context, problem,* and *solution* are the primary components of the pattern. Patterns provide reusable experienced solutions rather than reusable program logic. In general, a design pattern has the following essential elements: Name, Problem, Context, Forces, Solution, Examples, Resulting Context, Rationale, Related Patterns, and Known Uses [1]. Patterns have been classified into three broad categories: the *Creational Pattern* deal with object creation in multiple situations, e.g., Factory Pattern, Builder Pattern, Singleton Pattern, etc.; the *Behavioral Pattern* handles the communication mechanism between the objects, e.g., Interpreter Pattern, Mediator Pattern, and Observer Pattern; etc.; and, The *Structural Pattern* eases the design by identifying a simple way to realize relationships between entities, e.g., Adapter Pattern, Bridge Pattern, etc. Schmidt [3] has noted a number of benefits gained from incorporating design patterns into the development process, including: enabling widespread reuse of software architecture designs, improved communication within and across development teams, facilitated training of new programmers, and transcending ways of thinking imposed by individual programming languages.

The Web Ontology Language (OWL)[4] is a knowledge representation framework built on top of RDF[5] semantics and has three variants: OWL Lite, OWL DL, and OWL Full. The OWL DL (Description Language) variant is the most powerful and popular framework chosen for developing ontologies, maintaining a balance between expressiveness and reasoning complexity. As ontologies are designed for encapsulating domain conceptualization, the knowledge engineers are very much concerned in encapsulating all of the domain vocabulary for implementation purpose and exploiting reasoning algorithm

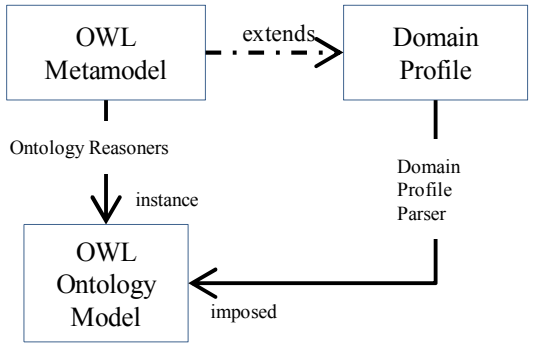


Figure 1: Architecture of OWL, OWL Domain Profile and OWL Ontology Model.

by using underlying description logic[6]. In our previous work[7, 13], modeling capabilities of ontologies were compared to standard modeling techniques such as the unified modeling language (UML), entity-relationship design (ERD), and the extensible markup language (XML)[7], with extensions were proposed to leverage modeling capabilities of OWL when compared to standard software modeling paradigms. One of the proposed features was the *OWL Domain Profile* (ODP) module, which supports OWL with metamodeling entities for capturing abstract domain specific concepts in a *domain profile* to act as metamodeling entities. The architecture of ODP in support with OWL and the OWL ontology is shown in the Figure 1, where the OWL is *extended* to develop ODP which is later imposed on the domain ontology. The *Domain Profile parser* (under development) authenticates and validates the imposing of the profile onto the model.

The concept of *domain profile* is illustrated in Figure 2,

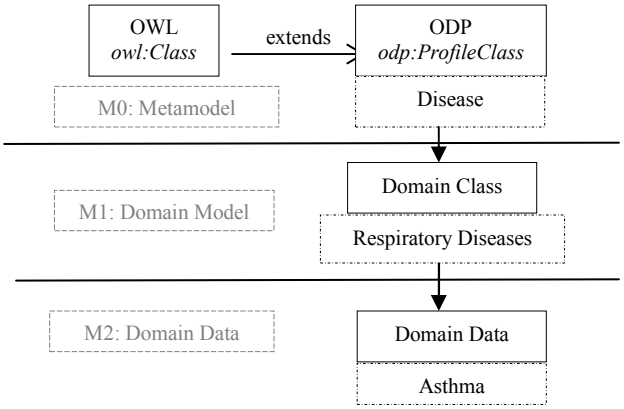


Figure 2: Layered representation of OWL with ODP.

where a domain model (M1 Level) provides the abstract view of the real world, while the metamodel (M0 Level) provides the abstract view of the domain model. For instance, in the domain of medicine (see Figure 2), the term *Disease* may represent all of the known diseases. The domain expert may further classify the concept *Disease* into various classes such as *Respiratory Diseases*, *Cardiac Diseases*, *Circulatory Diseases,* etc., which may be further classified forming a classic "is-a" hierarchical relationship between the classes. Similarly, concepts such as *Symptom, Treatments, Procedure, Surgery, Drugs*, etc., can be built to form complex medical ontologies. After defining the class model (at M1 Level), viewing from the perspective of collecting real world data (M2 Level), the concept hierarchies and associations between them provide an excellent schema for describing the data. However, from the perspective of the metamodel (M1 Level), concepts such as *Disease, Symptom, Treatment, Procedure, Drug*, etc. are domain specific generic concepts which can be abstracted and moved to metamodel level. Once these domain specific generic concepts are captured and connected they form a *semantic pattern* which can be reused in multiple domain

model settings irrespective of the domain model. The ODP framework used for representing these patterns at the metamodel level has the following entities: **odp:ProfileClass** a metamodel entity used to represent a profile class, **odp:Attribute** used to capture the characteristics of a profile class, **odp:ProfileDatatypeProperty** a metamodel entity used to capture interactions between the profile classes, and odp:ProfileDatatypeProperty represent datatype properties of a profile class. These entities are derived by extending OWL primitive entities [7]. In Figure 2, Disease, Symptom, Treatment, Procedure and Drug are of type *ProfileClass* and their interactions will be of type *ProfileObjectProperty*. Once defined, the profile entities (*Disease)* are *imposed* [2] onto the domain model entity (*Respiratory Disease* - M1) which is read as Respiratory Diseases *isofType* Disease. Later, the domain model classes can be instantiated to collect instance data (M2) such as *Asthma*. Similarly, the profileClass *Disease* can be imposed onto multiple domain classes such as "*Cardiac Diseases*", "*Respiratory Diseases*", "*Reproductive Diseases*", etc., are all of *type* "*Disease*".

Finally, to complete this section, we focus on prior and ongoing work in the area of design patterns as applied to ontologies and knowledge representation. Gangemi[8] has proposed a Conceptual Ontology Design Pattern (CODeP) to capture generalized use case scenario acting as a template to solve domain modeling issues. The CODeP are

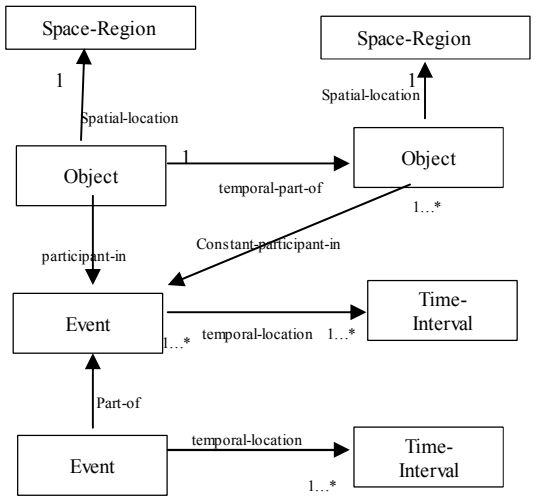


Figure 3: CODeP Participation Pattern.

developed by extracting key backbone knowledge on which the core ontology is built and is encoded in any knowledge representational format. For implementation, the author has defined a number of fundamental core patterns such as the participation pattern in Figure 3 which is extracted from the DOLCE ontology which illustrates participation relation between objects and events; the Role-Task pattern is also extracted from extended version of DOLCE explains the relationship between role, task, object and event. These patterns are implemented using the OWL framework by defining concepts such as Space-Region, Event, Object, Time-Interva,l etc., as OWL classes and associations such as Spatial-Location, temporal-part-of, part-of, participant-in, etc., using the OWL objectProperty. Our observation on this work is that the patterns developed by Gangemi(Figure 3) are captured as domain model (M1 in Figure 2), but not at the metamodel level (M0 in Figure 2). Staab[9] defined semantic patterns as a means of representing *epistemological* level concepts which can be instantiated by any target language. RDF is chosen for pattern development as RDF is basic building block for developing any other knowledge representational frameworks. A *Consistency Translation* is performed to assure semantics of the implemented language and pattern are the same. Clark [2] introduced the concept of "knowledge Patterns" which are defined as structure representing reoccurring pattern similar to software patterns, but morphing the knowledge pattern entities onto domain classes instead of instantiating them as show in Figure 4. In our approach to be presented in Section 3, a design pattern is defined as first-order logic which is later incorporated into the domain knowledge with definitive axioms. For example, a simple distribution design pattern as shown in Figure 4 where ***P*** is the producer, ***S*** is a switch mechanism, and ***C*** is the consumer. This pattern can be applied to any domain

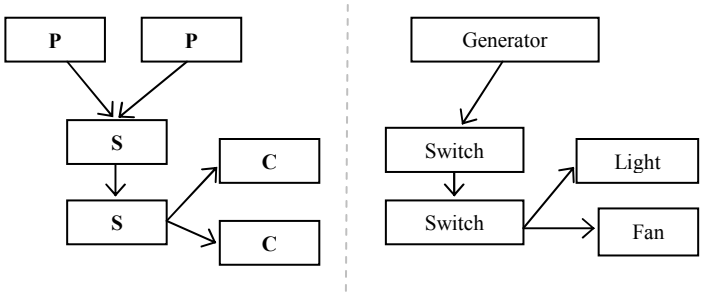


Figure 4: Distribution network pattern by Clark.

model, for instance, ***P*** can represent a *battery, power plant*, *reactor* etc., ***I*** can be a *common switch* and ***C*** can be any device which consumes power such as *light, heater, computer* etc. as shown in Figure 4.

3 Semantic Patterns using ODP

In this section, we propose the inclusion of semantic design patterns using the ODP feature and exemplify its development through the Ontology Meta Vocabulary (OMV) [10] model. A *Semantic Design Pattern* captures reoccurring structure along with it associated semantics at the metamodel level allowing ontology designers to solve similar problems occurring in multiple settings. As discussed in Section 2, the ODP entities capture generic abstract domain concepts which can be class, attribute, or relationship. Our proposal in this section involves, interrelating profile classes (type **odp:ProfileClass**) profile objectProperty (**odp:ProfileDatatypeProperty**) which will result in a semantic pattern which can be imposed to develop an ontology schema. As the concepts are captured at metamodel

level (M0 in Figure 1) a single semantic design pattern can be leveraged to develop multiple domain models. In order to demonstrate the development of semantic design pattern using ODP, we utilize the OMV model. OMV is a domain model for capturing metadata about ontology; Figure 4 shows the core entities of OMV.

As given in Figure 5, in a typical ontology engineering process, a *Person(s)* or *Organization(s)* is responsible for developing an ontology represented using *Ontology* entity. *OntologyTask* describes the primary task of the ontology and *LicenceModel* describes the usage boundaries of the ontology. The entities *OntologySyntax*/*OntologyLanguage* capture the implementation details of the ontology schema. A further classification of the ontology relies on their level of formality (in ontology, it can be defined as the depth of knowledge captured and its purpose [14]) captured using *FormalityLevel* and details on ontology languages are representable with the help of the classes *OntologySyntax*, *OntologyLanguage* and *KnowledgeRepresentationParadigm*. The domain the ontology describes is represented by the entity *OntologyDomain*, while *OnotologyType* describes the nature of the content of the ontology.

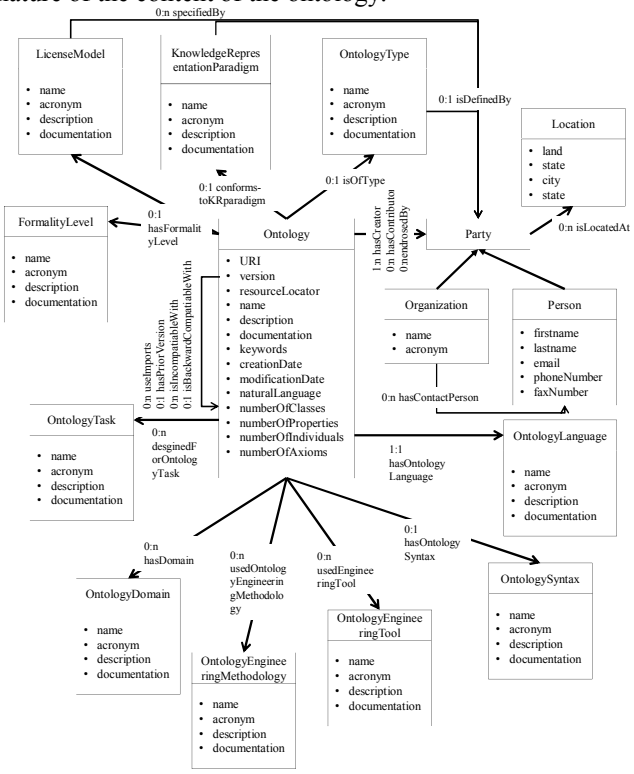


Figure 5: Overview of OMV Core concepts

Since OMV is a generic model to capture metadata of the ontology irrespective of the domain, we can transform OMV model entities into ODP metamodeling entities which can be connected to form OMV semantic design patterns. The ontology designer can impose the metamodel entities according to his/her usage and the perspective. The first step is to convert the OMV model classes, attributes, and associations into respective ODP entities forming the basic building block of the pattern.

In the first part of our proposal is to create and include semantic design patterns, the OMV classes (rectangles in Figure 4) are converted into ODP Profile class entities as shown in Table 1. In the Table 1, the left column represents the OMV classes and the right column represents equivalent ODP profile classes. For example, "Party" (second row in Table 1) is an OMV domain model class, while "rdf:ID=Party" is its ODP Profile class representation.

Table 1: OMV Classes to ODP ProfileClass.

OMV Model Class	ODP Profile Class
Party	<odp:ProfileClass rdf:ID ="Party"/>
OntologyDomain	<odp:ProfileClass rdf:ID "OntologyDomain"/>
OntologyTask	<odp:ProfileClass rdf:ID ="OntologyTask"/>
Ontology Engineering Methodology	<odp:ProfileClass rdf:ID ="OntologyEngineeringMethodology"/>
OntologySyntax	<odp:ProfileClass rdf:ID ="OntologySyntax"/>
LicenseModel	<odp:ProfileClass rdf:ID ="LicenseModel"/>
Location	<odp:ProfileClass rdf:ID ="Location"/>
FormalityLevel	<odp:ProfileClass rdf:ID ="FormalityLevel"/>
Ontology EngineeringTool	<odp:ProfileClass rdf:ID ="OntologyEngineeringTool"/>
OntologyLanguage	<odp:ProfileClass rdf:ID ="OntologyLanguage"/>
Organization	<odp:ProfileClass rdf:ID ="Organization"/>
Knowledge Representation Paradigm	<odp:ProfileClass rdf:ID ="KnowledgeRepresentationParadigm"/ >

Second, the OMV attributes in Figure 4 are mapped to the ODP profile datatypeProperty, if the range of the attribute is a datatype (e.g., integer, double, URL, string, etc.) as shown in Table 2. The values on the left column in Table 2 represents attributes of OMV domain classes and the right column represents the respective ODP representation. For example, "name" (second row in Table 2) is an attribute of OMV Domain class while its representation in ODP is of type ProfileDatatypeProperty (i.e., type of the attribute is String).

Table 2: OMV Attributes to ProfileDatatypeProperty.

OMV Attributes	ODP Profile DatatypeProperty
name	<odp: ProfileDatatypeProperty rdf:ID ="name"/>
acronym	<odp: ProfileDatatypeProperty rdf:ID ="acronym"/>
description	<odp:ProfileDatatypeProperty rdf:ID ="description"/>
documentation	<odp:ProfileDatatypeProperty rdf:ID ="documentation"/>
land	<odp:ProfileDatatypeProperty rdf:ID

	=”land”/>
state	<odp:ProfileDatatypeProperty rdf:ID =”state”/>
city	<odp:ProfileDatatypeProperty rdf:ID =”city”/>
firstName	<odp:ProfileDatatypeProperty rdf:ID =”firstName”/>
lastName	<odp:ProfileDatatypeProperty rdf:ID =”lastName”/>
email	<odp:ProfileDatatypeProperty rdf:ID =”email”/>
phoneNumber	<odp:ProfileDatatypeProperty rdf:ID =”phoneNumber”/>
faxNumber	<odp:ProfileDatatypeProperty rdf:ID =”faxNumber”/>
URI	<odp:ProfileDatatypeProperty rdf:ID =”URI”/>
version	<odp:ProfileDatatypeProperty rdf:ID =”version”/>
resourceLocator	<odp:ProfileDatatypeProperty rdf:ID =”resourceLocator”/>
keywords	<odp:ProfileDatatypeProperty rdf:ID =”keywords”/>
creationDate	<odp:ProfileDatatypeProperty rdf:ID =”creationDate”/>
modificationDate	<odp:ProfileDatatypeProperty rdf:ID =” modificationDate”/>
naturalLangauge	<odp:ProfileDatatypeProperty rdf:ID =” naturalLangauge”/>
numberofClasses	<odp:ProfileDatatypeProperty rdf:ID =” numberOfClasses”/>
numberOf Properties	<odp:ProfileDatatypeProperty rdf:ID =” numberOfProperties”/>
numberOf Individuals	<odp:ProfileDatatypeProperty rdf:ID =” numberOfIndividuals”/>
numberOfaxioms	<odp:ProfileDatatypeProperty rdf:ID =” numberOfAxioms”/>

Third, the OMV associations (interactions between classes in Figure 4) are mapped to ODP ProfileObjectProperty as shown in Table 3. The left column represents associations in the OMV domain model while the right column represents the respective ODP conversions are of type ProfileObjectProperty. In Table 3, “hasDomain” is association between classes “Ontology” and “Domain” (Figure 4) and “rdf:ID=hasDomain” is its representation ODP.

Table 3: OMV Associations to ODP ProfileObjectProperty.

OMV Associations	**ODP ProfileObjectProperty**
hasDomain	<odp:ProfileDatatypeProperty rdf:ID =” hasDomain”/>
usedOntology Engineering Methodology	rdf:ID =” usedOntologyEngineering Methodology”/>
usedEngineeringTool	<odp:ProfileDatatypeProperty rdf:ID =” usedOntologyEngineering Tool”/>
hasOntology Syntax	<odp:ProfileDatatypeProperty rdf:ID =” hasOntologySynatx”/>
hasOntology Language	<odp:ProfileDatatypeProperty rdf:ID =” hasOntologyLanguage”/>
hasContactPerson	<odp:ProfileDatatypeProperty rdf:ID =” hasContactPerson”/>
isLocatedAt	<odp:ProfileDatatypeProperty rdf:ID =” isLocatedAt”/>
hasCreator	<odp:ProfileDatatypeProperty rdf:ID =” hasCreator”/>
hasContributed	<odp:ProfileDatatypeProperty rdf:ID =” hasContributed”/>
endrosedBy	<odp:ProfileDatatypeProperty rdf:ID =” endrosedBy”/>
isOfType	<odp:ProfileDatatypeProperty rdf:ID =” isOfType”/>
conformstoKR paradigm	<odp:ProfileDatatypeProperty rdf:ID =” conformstoKRparadigm”/>
hasFormality Level	<odp:ProfileDatatypeProperty rdf:ID =” hasFormalityLevel”/>
useImports	<odp:ProfileDatatypeProperty rdf:ID =” useImports”/>
hasPriorVersion	<odp:ProfileDatatypeProperty rdf:ID =” hasPriorVersion”/>
isIncompatiableWith	<odp:ProfileDatatypeProperty rdf:ID =” isIncompatiableWith”/>
isBackward CompatiableWith	<odp:ProfileDatatypeProperty rdf:ID =” isBackwardCompatiable With”/>
desginedFor OntologyTask	<odp:ProfileDatatypeProperty rdf:ID =” desginedForOntologyTask”/>

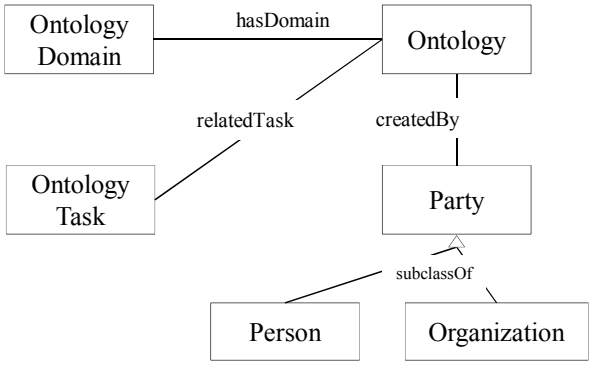


Figure 6: A Part of OMV Semantic Design Pattern

Once the OMV ODP entities have been defined, the next step in our proposed approach for semantic design patterns defines: ODP profile ObjectProperty entities which represent associations that can be used for interconnecting various ODP profile classes; and, ODP profile DatatypeProperty entities that can be associated as attributes to ODP profile classes. The ODP entities defined in Table 1-3 can be

- Acute nasopharyngitis (common cold)
- Acute sinusitis
- Acute pharyngitis
  - Streptococcal pharyngitis
  - Acute pharyngitis due to other specified organisms
  - Acute pharyngitis, unspecified
- Acute tonsillitis
- Acute laryngitis and tracheitis
  - Acute laryngitis
  - Acute tracheitis
  - Acute laryngotracheitis
- Acute obstructive laryngitis (croup) and epiglottitis
  - Acute obstructive laryngitis (croup)
  - Acute epiglottitis
- Acute upper respiratory infections of multiple and unspecified sites

Figure 7: ICD classification of *Acute upper respiratory infections*.

interconnected to form a semantic design pattern. To illustrate, Figure 6 shows a portion of an OMV semantic design pattern and essentially denotes that an ontology (represented by the entity *Ontology)* represents a domain (represented by the entity *OntologyDomain)*, is developed to accomplish a task (represented by the entity *OntologyTask),* and is created by an company (represented by the entity *Organization)* or person (represented by the entity *Person)*.

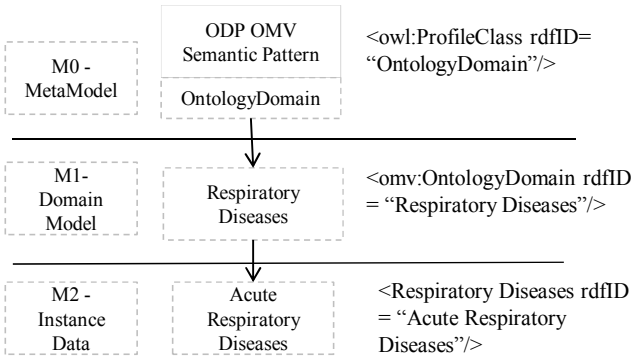


Figure 8: Capturing domain related information of the sample ontology shown in Figure 7.

```
<odp:ProfileClass rdfID="Organization"/>
<omv:Organization rdfID= "WHO"/>
<WHO rdf:ID=" World Health Orgnization- ICD
                                    10">
<odp:ProfileClass rdfID="Location"/>
<omv: Location rdfID= "Switzerland"/>
< Switzerland rdf:ID=" Geneva">
```

Figure 9: Syntax for capturing ontology information.

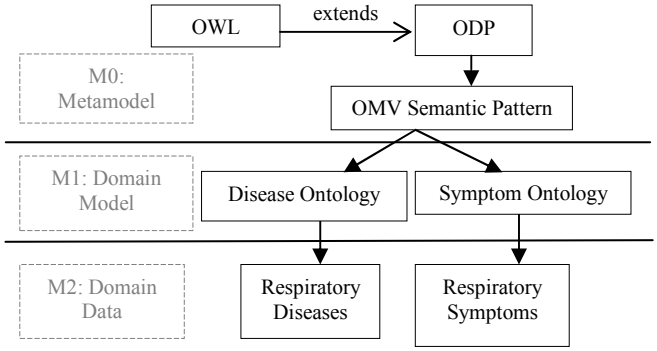


Figure 10: Applying OMV semantic Pattern on multiple ontologies.

The pattern entities *Ontology, Domain, Task, Party, Person* and *Organization* are of type ProfileClass; association's *hasDomain, relatedTask*, *createdBy* and *subClassOf* are of type ProfileObjectProperty. In order to apply the semantic design pattern, consider the ontology in Figure 7, a snapshot taken from the ICD[11] taxonomy which classifies various "*Acute Upper Respiratory Diseases*". For capturing the domain related information for the sample ontology, the **OntologyDomain** is instantiated as profileClass. This entity **OntologyDomain** at the M0 Level is imposed to capture domain information (Respiratory Diseases *type* OntologyDomain) of the ontology model at the M1 level. Then, the ontology model entity "Repository Diseases" at M1 level can be instantiated to capture the instance data about the ontology (Acute Upper Respiratory Diseases in Figure 8) at the M2 level. Other metadata on the ontology (Figure 9) such as *organization*, *location* etc. can be captured using the OMV semantic pattern design entities **Organization**, and **Location** respectively using the syntax given in Figure 9. As the semantic design pattern and its entities are defined at the metamodel level, they can be reused to describe multiple ontology models. Shown in Figure 10, the OMV semantic design pattern is defined using ODP module at the M0 Level. The same pattern is reused to capture information about multiple ontology models such as *"Disease Ontology"*, *"Symptom Ontology",* and *"Medication Ontology"* at the M1 level. The ontology models at the M1 level are later instantiated to capture instance data such as "Respiratory Diseases", "Respiratory Symptoms" etc.

4 Conclusion and Ongoing Work

Design patterns in software engineering are templates that represent the repetitive structure occurring problem solutions that are similar across multiple domains. In this paper, we have proposed the use of semantic design patterns to leverage these capabilities for ontology design and development. We believe we have demonstrated that these reusable structures and their semantics in the form of design patterns can provide a benefit that enhances both the

semantic and the reusability of ontologies. Overall, this paper proposed an approach to model semantic design patterns using the OWL Web Ontology Language Domain Profile feature that integrated the design pattern concept directly into the domain ontology. Towards this objective, in Section 2 we provided background and motivation on design patterns, an OWL ODP extension, and research efforts that apply design patterns to ontologies and knowledge engineering. Using this as a basis, Section 3 proposed the semantic design patterns using the OMV model and demonstrated that capturing the pattern entities at the metamodel level reuses these components in multiple setting. Our ongoing work has utilized the UML meta-model process [7] to provide a more software engineering like process to ontology design, development, and deployment; the inclusion of semantic design patterns is an extension of this work to enhance the modeling process. For implementing the ODP, we are considering open source ontology editors such as Protégé, JOE, Hozo etc. which support OWL/RDF frameworks.

5 References

[1] E. Gamma, R. Helm, R. Johnson and J. Vlissides, "Design Patterns: Elements of Reusable Object-Oriented Software", Addison-Wesley Professional, 1st ed., 1994.

[2] P. Clark, J. Thompson, and B. Porter, "Knowledge Patterns", Handbook of Ontologies, pp. 191-207, Springer, 2003.

[3] D. C. Schmidt, "Experience Using Design Patterns to Develop Reusable Object-Oriented Communication Software", Special Issue on Object-Oriented Experiences, Vol.38, 1995.

[4] L. Lacy, "Owl: Representing Information Using the Web Ontology Language", Trafford Publishing, 2005.

[5] S. Powers, "Practical RDF", 1st ed., O'Reilly Media, 2003.

[6] W. Kuhn, "Modeling vs. Encoding for semantic web," Semantic Web Journal, Vol. 1, 2010.

[7] R. Saripalle, S. Demurjian, and S. Berhe, "Towards Software Design Process for Ontologies", Intl. Conf. on Software and Intelligent Information, October, 2011.

[8] A. Gangemi, "Ontology Patterns for Semantic Web Content", Proc. Of 4th Intl. Semantic Web Conf., pp. 262-276, 2005.

[9] S. Staab, M. Erdmann, and A. Maedche, "Engineering Ontologies using Semantic Patterns", Proc. Of Intl. Joint Conf. on Artifical Intelligence, 2001.

[10] J. Hartmann, R. Palma, Y. Sue, P. Hasse, and M. Suarez-Figueroa, "OMV- Ontology Metadata Vocabulary", Proc. of Intl. workshop on Ontology Patterns for the Semantic Web, November, 2005.

[11] International Classification of Diseases, http://www.who.int/classifications/icd/en/, 2009.

[12] Wim De Pauw, David Lorenz, John Vlissides, and Mark Wegman, "Execution Patterns in Object-Oriented Visualization", Proc. Of 4th Conf. on Object-Oriented Technologies and Systems, 1998.

[13] S. Demurjian, R. Saripalle and S. Berhe, "An Integrated Ontology Framework for Health Information Exchange", Intl. Conf. on Software Engineering and Knowledge Engineering, August, 2009.

[14] N. Guarino, "Formal Ontology in Information Systems", Proc. of FOIS, Vol. 1, pp. 3-15, June 1998.

An ontology approach to enhance interoperability for musculoskeletal problems

A.Tara Sampalli[1] and B. Mary Lynch[2]
[1]Medical Informatics, Dalhousie University, Halifax, Nova Scotia, Canada
[2] Anesthesia and Pharmacology, Dalhousie University, Halifax, Nova Scotia, Canada

Abstract *- Musculoskeletal problems are leading causes of disability in adults with high health related costs to the system. Evidence suggests that obscure and inconsistent domain knowledge can create challenges to timely and relevant communication among multidisciplinary care providers in the management of these conditions. The primary objective of this study was to develop, test and evaluate a model and a methodology for creating an ontology in the heterogeneous domain of musculoskeletal and pain related problems. The methodology applied a two-staged approach for enabling interoperability, namely, development of a controlled vocabulary followed by an ontology. The ontology was developed from the knowledge that existed in 100 patient charts and among 8 domain experts. The chronic pain ontology contained 182 classes and over 51 data and 38 object properties. Sixty-seven percent of clinicians agreed on the overall usefulness of the ontology as a boundary object. Goal of this research was to enable better communication among multidisciplinary care providers through an ontology based interoperability.*

Keywords: Ontologies, musculoskeletal problems, interoperability, multidisciplinary.

1 Introduction

Multidisciplinary care teams have come to the forefront as an effective management strategy for musculoskeletal problems [1-4]. Chronic low back pain (CLBP) is a complex condition requiring timely, cost effective and multidisciplinary care management and is significant in terms of related healthcare costs in Canada and the rest of the world [3,4]. CLBP is usually caused by a trauma to the lower back or certain diseases such as, arthritis. Acute back pain is typically managed by family physicians. However, individuals with chronic back pain are more complex in their needs and require multidisciplinary care management [1,2]. Studies have shown the consequences of poor communication among multidisciplinary care providers resulting in poor care experiences, over prescription of medical tests, misdiagnoses, delayed care and inaccurate treatment plans [5-7]. Studies have also demonstrated that clinical documentation is the primary source of communication in multidisciplinary management [8-11]. Studies have discussed the significant need to identify methods to improve collaboration and communication among care providers in order to facilitate seamless care for patients with complex health conditions [8].

In homogeneous environments, there are many approaches that have been discussed as methods to improve collaboration in health care such as controlled vocabulary, ontology, through health care routines, and medical rounds [12, 13].

There are many challenges to effective collaboration among multiple disciplines in the management of complex health conditions. Barriers to effective communication include dealing with poorly categorized knowledge with multidisciplinary, inconsistent and non-standardized clinical documentation with new knowledge emerging among various knowledge communities or groups of experts.

Figure 1 shows these important challenges to communication among care providers in the management of complex chronic conditions.

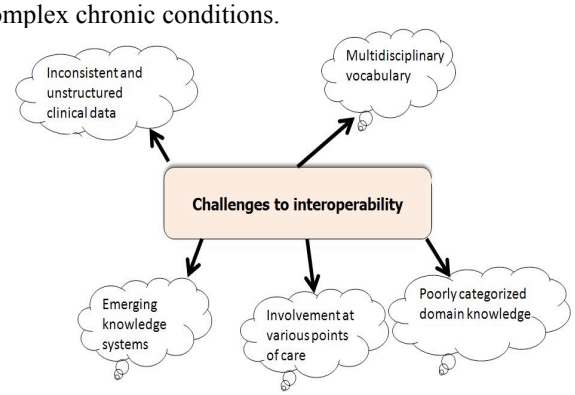


Figure 1: Interoperability in heterogeneous domains

CLBP and many other musculoskeletal problems have many of the outlined challenges to communication. Knowledge management and organization in heterogeneous domains of such complex conditions is a grand challenge for health informatics. In this paper, heterogeneous domains are defined as domains that are poorly categorized, multidisciplinary, non-standardized and inconsistent.

Ontologies have gained importance in recent years as a knowledge management platform in many areas including health care [13-15]. An ontology is "an explicit and formal specification of a conceptualization" [16]. An ontology consists of a finite list of terms or concepts and the relationship between these terms. Ontologies are preferred to conventional classifications due to the higher level of expressiveness that is possible in describing concepts and their relationships [14]. Despite their high level of

specification of these classes and relationships, ontologies also allow a great deal of flexibility. Ontologies have been typically developed in stable knowledge domains [17,18].

2 Related Work

Challenges in developing ontologies in heterogeneous domains have been discussed in the literature.

Dominigue et al. [14] identify the key requirement for an ontology approach to knowledge management as a community's perspectives being stable on an issue with "well defined roles", "specified criteria" and "codified procedures". Challenges related to developing ontologies when there is a lack of consensus in a community are discussed in the subsequent paragraphs.

A study by Larson and Martone [17], the challenges of formalizing knowledge for neuroscience were explored. The authors claimed that formalizing knowledge about poorly understood biological systems presents many obstacles to the development of ontologies. This study highlighted the importance of developing a layer of standardization prior to attempting higher level specification such as the creation of ontologies in the domain.

In a study by Lin et al. [18], the challenges of a mental health group of professionals working with emerging knowledge was discussed. This study describes the challenges and importance of building knowledge through ontologies in heterogeneous situations. This study presented the preliminary challenges that exist in the knowledge capture for a domain that has obscure definitions, lack of consensus, unstructured data, inconsistent use of vocabulary and assessment scales. A significant challenge encountered in this work was to bring structure to knowledge that continues to be generated in an ad hoc manner.

In a study by Qin and Paling [19], the importance of developing ontologies in heterogeneous domain was examined. The research describes the creation of an ontology from a well defined and well used controlled vocabulary in order to provide a higher level of semantics to the concepts in the vocabulary. Digital objects, such as those in the Gateway to Educational Materials (GEM ontology) encompass multiple dimensions of characteristics which often play important roles for users in search of precise information in an efficient manner. The authors suggest that a conventional cataloguing code will be inadequate to describe these details in a lesson plan, as many of these elements do not even exist in the vocabulary. In this study, the authors developed an ontology with the intention of adding another layer of semantic operability to the terminologies found in controlled vocabularies.

In this paper, we discuss the development of a model and methodology to enable interoperability through an ontology in the heterogeneous domain of musculoskeletal problems such as chronic low back pain.

3 Proposed Model for Interoperability

Figure 2 illustrates the architecture of our proposed two-staged approach to develop interoperability in the heterogeneous domains of complex conditions.

The first stage was the development of a controlled vocabulary to enable standardization, organization and consistency of the heterogeneous domain knowledge [20]. The second stage was the development of an ontology from the controlled vocabulary to enable formal representation of the domain concepts, description of the domain concepts and specification of relations between the concepts [19]. The interoperability developed have the characteristics at the pragmatic level (knowledge translation) of shareability, have the capacity to be dynamic in nature and are in standardized forms.

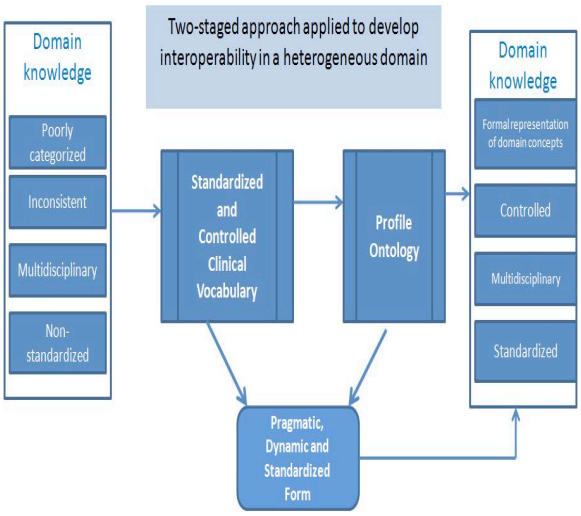


Figure 2: Architecture of the model and methodology

The first stage of the two-staged approach included the creation of a standardized and controlled clinical vocabulary. SNOMED CT® [21], a widely used reference terminology was used to standardize the concepts and terminologies found in the patient charts. A pragmatic approach [20] was applied in the development of the controlled vocabulary as the domain knowledge is heterogeneous.

The method for creating the controlled vocabulary was driven by the purpose of generating the goal and usage of the vocabulary The development of the controlled vocabulary involved the creation of standardized and controlled clinical vocabularies at the levels of syntactic, semantic and pragmatic interoperability: chart audit and interviews with experts to identify key concepts in the domain of the complex condition (syntactic), standardization of the vocabulary by establishing concrete meaning for concepts (semantic), and testing and evaluation of the vocabulary by the users to evaluate the potential for knowledge translation (pragmatic). The chart audit and interviews with experts helped generate the vocabulary. SNOMED CT® was used as a reference

terminology to standardize the terms retrieved in the chart audit process. The re-coding of patient profiles, evaluation and feedback from the domain experts tested and evaluated the vocabulary. A further step in the evaluation included feedback from clinicians in the community.

The second stage of the two-staged approach was the creation of an ontology in the heterogeneous domain consisting of 3 phases: Development, testing and evaluation. The development phase included the experts in the domain specifying and organizing the knowledge in the domain. This phase primarily drew the knowledge from the controlled vocabulary. The testing phase included the clinicians browsing the profile ontology developed in this research to examine the concepts in the ontology, the relationships between concepts, concept attributes and the individuals populated in the ontology. Following this was an evaluation phase that included feedback from the domain experts on the overall usefulness of the ontology in patient care with emphasis on usefulness from a health discipline perspective, from other health disciplines and the multidisciplinary nature of interactions captured in the ontology.

Protégé 3.4.2 was used to implement the patient profile ontology [22]. The profile ontology was exported into the Web Ontology Language (OWL). A consistency check of the classes in the ontology was conducted. Consistency checking helped detect classes that cannot have instances.

The implementation phase also included the evaluation of the ontology by domain experts for accuracy, completeness and usefulness of the knowledge represented in the ontology. The evaluation phase included the clinicians browsing the ontology using an ontology browser. They browsed various aspects of the ontology such as the classification scheme, multidisciplinary relations between concepts, instances, and standardization of concepts. Google ontology browser was used by clinicians to browse the ontology [23]. They provided feedback on the usefulness of the ontology through a survey questionnaire. Specifically, they offered feedback on the overall usefulness of the ontology, the relevance of the ontology in the context of patient care and the value of shared knowledge in the multidisciplinary domain. Individuals or instances are used in the profile ontology to present list of concrete concepts of relevance for each class.

4 Results

A complex and chronic health condition, namely, chronic pain was selected to test the viability of the proposed methodology in heterogeneous knowledge systems. One-hundred patients, 8 domain experts and 42 multidisciplinary community clinicians participated in the development of the chronic pain vocabulary and ontology.

4.1 Profile ontologies for chronic pain and musculoskeletal problems

The ontologies present a detailed taxonomic overview of the domain of complex health conditions.

The profile ontology for chronic pain contained 345 classes describing the profile concepts for the condition of chronic pain. At the basic level there are three relevant super-classes under the primary areas of health focus identified for the condition of chronic: *Medical, Physical* and *Psychosocial* as shown in Figure 3. The profile ontology includes definitions of over 80 properties, with 51 data and 38 object properties.

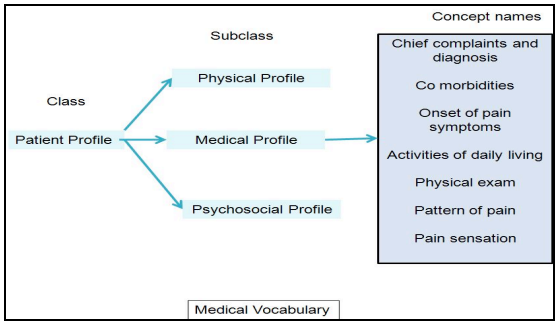


Figure 3: Vocabulary for chronic pain

The profile ontologies contained explication of all concepts included in the ontology such as the multidisciplinary nature of patient profile, the management scheme and the various concepts under each area of health focus. The properties in the ontologies introduce relations among concepts. A patient *HasOrganization* and the organization are inversely linked to the class Patient by *HasPatient*. The class Profile is linked to the class Management Scheme by property *hasCollaborativeManagement*. The class Psychosocial Profile is linked to the management scheme by property *ManagementRequired* which has individual *dietitian_referral* or *physician_referral*.

Standardized concepts are specified with their SNOMED CT ID number (Concept Unique Identifier) and with a list of synonyms. In the chronic pain ontology, standardized concepts are specified with their SNOMED CT ID number (Concept Unique Identifier) and with a list of synonyms. Class *Lumbar spine - tender* has a SNOMED CT® concept ID of 298673002 with parent concept being *Finding of sensation of lumbar spine* with finding site as lumbar spine structure.

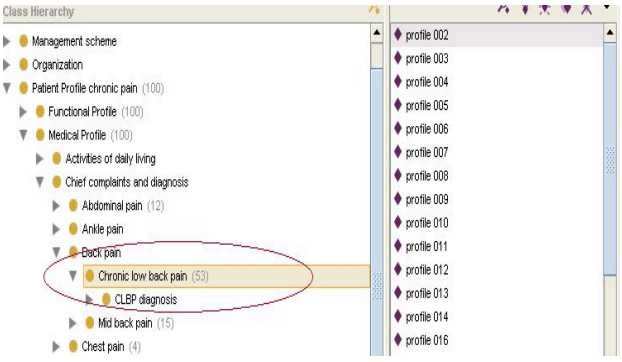


Figure 4: Query of "chronic low back pain" in the ontology

Figure 4 shows the query of "chronic low back pain" in the ontology retrieving that 53 patients have this diagnosis in the ontology.

Query of a symptom such as *Lumbar spine – tender* shows the number of patients with the symptoms and the super class of the concept in the ontology. The instances in profiles show the multifaceted nature of symptoms as substantiated under each area of health focus that exist in the domain of a patient.

Pain symptom as presented in the patient charts has been viewed in the patient charts by a psychotherapist, physician or physiotherapist from various angles of importance such as pattern of pain, anatomical site or in relation to the pain threshold. Figure 4 shows the view of a patient profile that shows the multidisciplinary care involved in the management of *Pain symptom* shown in Figure 5.

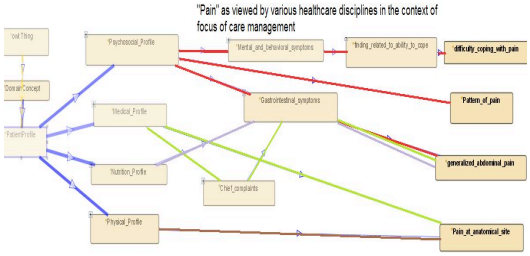


Figure 5: Multidisciplinary interactions in the categorization of Pain.

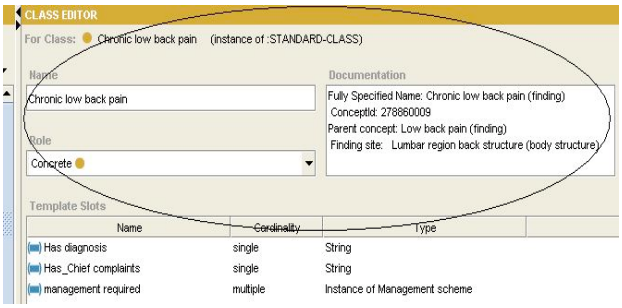


Figure 6: Standardized knowledge in the ontology.

Figure 6 shows another view in the ontology where the standardized information on each concept has been compiled using SNOMED CT.

Challenges in developing ontologies in heterogeneous domains have been discussed in the literature.

Dominigue et al. [14] identify the key requirement for an ontology approach to knowledge management as a community's perspectives being stable on an issue with "well defined roles", "specified criteria" and "codified procedures". Challenges related to developing ontologies when there is a lack of consensus in a community are discussed in the subsequent paragraphs.

A study by Larson and Martone [17], the challenges of formalizing knowledge for neuroscience were explored. The authors claimed that formalizing knowledge about poorly understood biological systems presents many obstacles to the development of ontologies. This study highlighted the importance of developing a layer of standardization prior to attempting higher level specification such as the creation of ontologies in the domain.

In a study by Lin et al. [18], the challenges of a mental health group of professionals working with emerging knowledge was discussed. This study describes the challenges and importance of building knowledge through ontologies in heterogeneous situations. This study presented the preliminary challenges that exist in the knowledge capture for a domain that has obscure definitions, lack of consensus, unstructured data, inconsistent use of vocabulary and assessment scales. A significant challenge encountered in this work was to bring structure to knowledge that continues to be generated in an ad hoc manner.

In a study by Qin and Paling [19], the importance of developing ontologies in heterogeneous domain was examined. The research describes the creation of an ontology from a well defined and well used controlled vocabulary in order to provide a higher level of semantics to the concepts in the vocabulary. Digital objects, such as those in the Gateway to Educational Materials (GEM ontology) encompass multiple dimensions of characteristics which often play important roles for users in search of precise information in an efficient manner. The authors suggest that a conventional cataloguing code will be inadequate to describe these details in a lesson plan, as many of these elements do not even exist in the vocabulary. In this study, the authors developed an ontology with the intention of adding another layer of semantic operability to the terminologies found in controlled vocabularies.

In this paper, we discuss the development of a model and methodology to enable interoperability through an ontology in the heterogeneous domain of musculoskeletal problems such as chronic low back pain.

4.2 Evaluation of the ontology

Community clinicians (n=42) and domain experts (n=8) reviewed the vocabulary for chronic pain. Only the domain experts (n=8) reviewed the ontology. Google ontology browser [23] was used by clinicians to browse the ontology and offer their evaluation as shown in Figure 7. They viewed the individual patient profiles, multidisciplinary information relevant to their discipline and other disciplines, in-depth query of symptoms and their management scheme.

The clinicians also viewed information on various symptoms including the profiles under which a symptom was categorized, the number of patients that had a symptom and the standardization information for the symptoms.

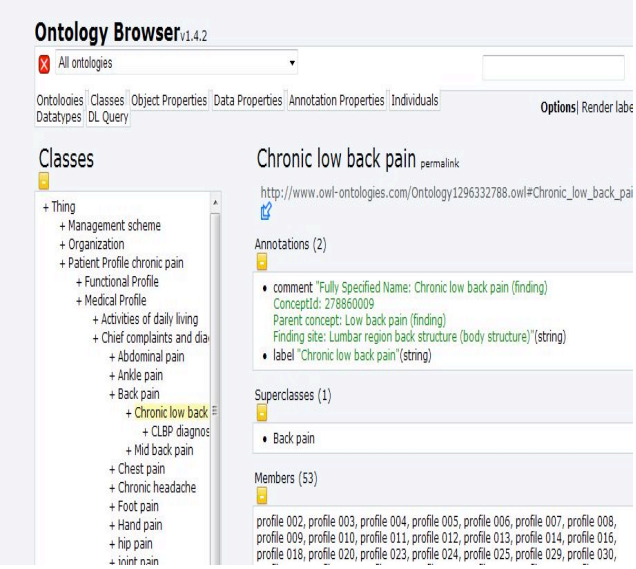


Figure 7: View of ontology using Google Ontology Browser.

Figure 8 shows the evaluation of the ontologies by the domain experts (n=8). Sixty-seven percent of clinicians agreed on the overall usefulness of the ontology as a boundary object. Highes level of agreement (83%) was reached on the usefulness of the ontology to view the information generated from other health disciplines in the categorization of a patient profile. The ontology had a consistently small to moderate percentage of clinicians showing strong agreement on its usefulness on all categories of the questionnaire. The ontology also had a very small percentage of disagreement on all categories of the survey questionnaire.

Cohen's Kappa and Kendall's Tau was used to determine level of agreement among the raters. Kendall's Tau was calculated at 0.6 with a moderate level of concordance among the 42 (multiple) raters with a p value of 0.03. Cohen's Kappa for the dietician's group showed the highest level of agreement with a score of 0.84.

5 Discussion and conclusions

A novel methodology and model has been presented in this research for the development of ontologies in heterogeneous knowledge domains. The broad objective of the research was to enhance communication in the multidisciplinary care management of chronic, complex and lesser known health conditions. The ontology approach was selected to develop consistency, standardization, organization and interoperability of domain knowledge with the broad goal of improving collaboration and communication for multidisciplinary clinicians involved in the care of patients with complex chronic conditions.

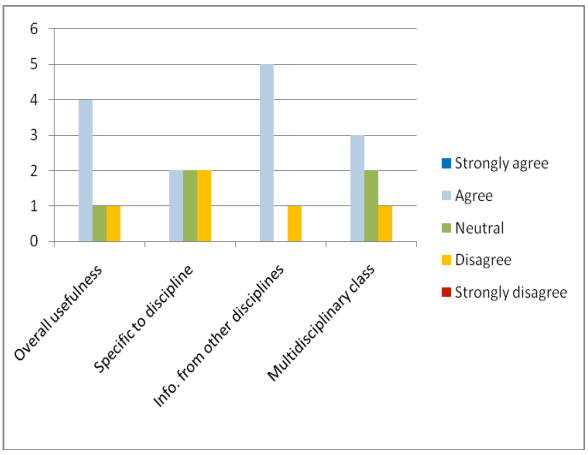


Figure 8: Evaluation of the ontology.

The development of the profile ontologies in this study was divided into three phases: specification, conceptualization and implementation [16]. The methodology includes several key components or criteria that were identified in past research such as acknowledging the heterogeneous nature of the domain knowledge [17] involving clinicians (experts and non-experts) in the process of development and evaluation and exploring the potential of the study by testing it in clinical workflow [18]. However there are several limitations to this research such as the scope being limited to the domain of patient profile information, a convenience sample of participants, size of the sample, the fact that the potential of the boundary objects in improving communication or collaboration among clinicians or the impact on patient care was not explored. The results do indicate that this direction of research has significant potential and requires further exploration.

0An ontology can reach a wider audience and has been deliberately selected to explicate the knowledge of lesser known and complex health conditions. Ontologies provide a pragmatic interoperable format for collaborative sharing of knowledge across communities of practice. The ontology has the potential to get richer as more users contribute new knowledge and as more patient instances are populated in the ontology. The overall agreement shown by experts in this study is very promising for the use of ontologies in the heterogeneous domains of complex health conditions.
Number in square brackets ("[ ]") should cite references to the literature in the main text. List the cited references in numerical order at the very end of your paper (under the heading `References'). Start each referenced paper on a new line (by its number in square brackets).

6 References

[1] Dysvik, E., Natvig, G. K., Eikeland, O. J., Brattberg, G., 2005. Results of a multidisciplinary pain management program: a 6- and 12-month follow-up study. Rehabilitation nursing: the official journal of the Association of Rehabilitation Nurses. Vol. 30, No. 5, 198-206.

[2] Peng, P., Stinson, J. N., Choiniere, M., Dion, D., Intrater, H., LeFort, S., Lynch, M., Ong, M., Rashiq, S., Tkachuk, G., Veillette, Y., 2008. Role of health care professionals in multidisciplinary pain treatment facilities in Canada. Pain Res. Manag. Nov-Dec: 13(6): 484-8.

[3] Cassidy JD, Côté P, Carroll LJ, Kristman V. Incidence and course of low back pain episodes in the general population. Spine. 2005; 30(24): 2817–2823.

[4] Brown A, Angus D, Chen S, Tang Z, Milne S, Pfaff J, Li H, Mensinkai S. Costs and outcomes of chiropractic treatment for low back pain [Technology report no 56]. Ottawa: Canadian Coordinating Office for Health Technology Assessment; 2005.

[5] Kennedy, 2008. Clinical documentation improvement in MS-DRGs as a strategy for compliance: facilities may consider clinical documentation audits to look for coding errors. Journal of Healthcare Compliance.

[6] Pace, W. D., Dickinson, L. M., Staton, E. W., 2004. Seasonal variation in diagnoses and visits to family physicians. Ann Fam Med, Vol.2, 411–7.

[7] Schoen, C., Osborn, R., How, S. K. H., Doty, M. M. and Peugh, J., 2008. In Chronic Condition: Experiences of Patients with Complex Health Care Needs, in Eight Countries, Health Affairs Web Exclusive, 1-16

[8] Stange KC, Nutting PA. Bursting the Bubble on Chronic Disease Management, the Meaning of Healing. Ann Fam Med. 2005;3(3):194–196.

[9] Ruhstaller T, Roe H, Thurlimann B, et al. The multidisciplinary meeting: An indispensable aid to communication between different specialties. European Journal of Cancer 2006;42:2459-62.

[10] Brooks, P. 2003. The impact of chronic illness: partnerships with other health care professionals, MJA 2003,179 (5), pp. 260-262

[11] Walsh. The clinician's perspective on electronic health records and how they can affect patient care. BMJ. 2004 May 15; 328(7449): 1184–1187.

[12] Lemieux-Charles, L. and W.L. McGuire. 2006. "What Do We Know about Health care Team Effectiveness? A Review of the Literature." Medical Care Research and Review 63: 263-300.

[13] Baneyx, A., Charlet, J., Jaulent, M., 2005. Building an ontology of pulmonary diseases with natural language processing tools using textual corpora. Studies In Health Technology And Informatics.

[14] Domingue, J., Motta, E., Shum, S. B., Vargas-Vera, M. and Kalfoglou, Y., 2001. Supporting ontology driven document enrichment within communities of practice. In: 1st International Conference on Knowledge Capture.

[15] Mostefai S, Bouras A and Batouche M., 2006. Effective collaboration in product development via a common sharable ontology. International journal of computational intelligence.

[16] Noy, N. F., and Mcguinness, D., 2000. Ontology development 101: A guide to creating your first ontology. Stanford KSL Technical Report KSL-01-05.

[17] Larson, S. D. and Martone, M. E., 2009. Ontologies for neuroscience: What are they and what are they good for? FrontNeurosci. May; 3(1):60

[18] Lin, Y., Poschen, M., Procter, R., 2006. Ontology as a social-technical process: a case study. ORA Conference/Workshop Paper.

[19] Qin, J., and Paling., 2001. Converting a controlled vocabulary into an ontology: the case of GEM. Information Research, 2001; 6(2)

[20] Carlile, P. R., 2002. A pragmatic view of knowledge and boundaries: Boundary Objects in new product development. Organization Science, 13 (4): 442-455.

[21] SNOMED CT® @ IHSTDO, 2008. SNOMED CT Clinical Terms Basics. Retrieved from http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/Recourses/Introducing_SNOMED_CT/SNOMED_CT_Basics_IHTSDO_Taping_Aug08.pdf

[22] Knublauch, H., 2004. "The Protégé OWL Plugin," in 7th International Protégé Conference, Bethesda, MD.

[23] Horridge, M., Drummond, N., Goodwin, J., Rector, A., Stevens, R. and Wang, H., 2006. OWL experiences and directions workshop. Retrieved from http://www.webont.org/owled/2006/acceptedLong/submission_9.pdf

[24] Verhaak, P. F., Meijer, S. A., Visser, A. P., Wolters, G., 2006. Persistent presentation of medically unexplained symptoms in general practice. Fam Pract. Vol. 23, 414–420.

How to make Ontologies self-building from Wiki-Texts

Bastian HAARMANN, Frederike GOTTSMANN, and Ulrich SCHADE
Fraunhofer Institute for Communication, Information Processing & Ergonomics
Neuenahrer Str. 20, 53343 Wachtberg, GERMANY

Abstract – While it used to be satisfactory that search applications used to look up simple keywords in the past, nowadays a search engine is increasingly expected to be able performing semantic search as well. This leads to an intensified research on ontologies supporting these applications. Users might only have a vague idea of the exact keywords to the content they seek, especially when it comes to technical or subject-specific topics and professional expertise is needed. A domain-specific ontology can support the user suggesting differentiated or alternative keywords for his query. Nevertheless, building an ontology containing expert knowledge is effortful and domain experts normally do not have sufficient skills how to build an ontology. Therefore, it is desirable to make an ontology automatically build and extend itself from already available information, such as wiki-texts. We developed a text mining component which automatically derives an ontology from collaborative texts written by scientific experts. Our component is used in a search application of an information system on energy research.

Keywords: Information Extraction, Ontologies, Semantic Search, Text Mining, Natural Language Processing

1 Introduction

The component we present in this paper is part of a project on energy research called "EnArgus" [1] which is currently under development[1]. The project's goal is to build an integrated system providing information on German energy research projects and funding to non-expert users, such as Parliament members, media representatives or the public. The information about projects funded by federal ministries is stored in different underlying databases. The system is supported by an energy-specific ontology which is automatically derived from a German Wikipedia-like text collection collaboratively written by energy research experts from universities, fellow Fraunhofer Institutes and a research center from the Helmholtz Association. The text collection serves on the one hand as a basis for the ontology to be derived. On the other hand, it will be used as a glossary for the non-expert users. Computational linguists created a text mining component using natural language processing.The component extracts the information from the wikis and builds resp. extends the resulting ontology. The next section explains the steps of the information extraction.

2 Text Mining & Information Extraction

The system's text mining component uses extensive linguistic rules adapted to the specific wording of the energy domain. The framework we use for information extraction is called GATE (General Architecture for Text Engineering; [2, 3]). The wiki-text collection is processed by a chain of several modules within GATE. Annotations are used in order to identify linguistic structures, such as syntax and semantics, as well as ontological knowledge structures such as classes, individuals, labels and properties. The process chain comprises the following modules:

Figure 1. The text mining component's chain of processing modules

2.1 Tokenizer

A tokenizer identifies individual tokens of the text. Tokens basically are single words. The tokenizer segments them by the spaces in between. Furthermore, tokens that are numbers, abbreviations and punctuation marks occurring in the text are recognized by the tokenizer as well. The resulting *Token* annotations contain meta information about the words.

2.2 Gazetteer

A Gazetteer is a look-up module which compares the obtained token annotations to specific lists of words. The Gazetteer lists contain different types of names and abbreviations. A matching word is labeled *Lookup* and receives the features of the respective type of the list, in which the matching word was found.

[1] This project was supported by the German Federal Ministry of Economics and Technology through grant number 03ET1064A. The content of this paper is solely the responsibility of the authors.

2.3 Sentence Splitter

The Sentence Splitter segments the text into whole sentences. Not every occurring full stop in the text forms a sentence end. Text elements such as "Mr.", "Dr." as well as other abbreviations, internet and IP addresses needs to be excluded from the splitting. The output annotation is *Sentence*. In a later step of the processing pipeline, the boundaries of subordinate clauses or extended participles and infinitives are detected and annotated accordingly.

2.4 Part-of-speech Tagger

The part-of-speech tagger (POS Tagger) determines the category of each word. It adds the feature category to the *Token* annotation and enters a value for the respective class (e.g. NN for normal noun). The POS tagger to be used depends on the desired strategy of tagging. There are older POS taggers using a rule-based tagging according to a lexicon and a rule set obtained from corpus work, such as the Hepple Tagger [4] which is derived from the Brill Tagger [5]. Newer approaches towards a more accurate POS tagging include tagging on the basis of statistics (TnT Tagger [6]) or neuronal networks such as Hidden Markov Models trained on large corpora (Stanford Tagger [7], Tree Tagger [8]). Empirical research shows that the latter strategy leads to a significant higher accuracy of the POS tags obtained, up to 97.5% [9]. The tagger we implemented makes use of a trained Hidden Markov Model which is able to assign each word both its category and its basic form (lemma), reduced from flexion morphemes for case, gender, number, etc.

2.5 General Transducer

In the next step of the process, the previous made annotations are used in order to identify syntactic and semantic information. Linguistic grammars stored in a module named General Transducer contain patterns for the detection of nominal phrases, prepositional phrases, verb groups and many more syntactic structures. We use this part of the General Transducer module as a chunker instead of a complete parser module. A parser would try to calculate a complete parse tree which represents the syntactic structure of each sentence. However, the complete and sometimes rather complex parse trees are normally not needed. It is often sufficient to know the verb group and the other constituents of a sentence as well as their sequence. Even more, parsing might fail because of unknown words or ambiguity while chunking is more robust in this matter. Additionally, parsing is on the one hand highly time-consuming and on the other hand computationally very resource-intensive. With the problems of parsing in mind, we substituted the parser by a chunker in the first part of the General Transducer. [10].

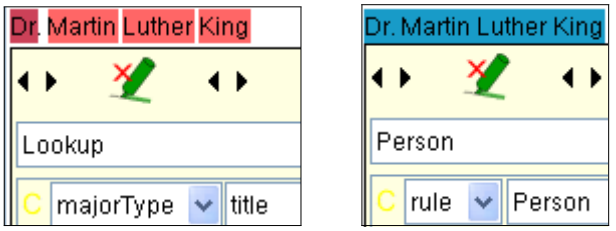


Figure 2. Named Entity Recognition within the General Transducer

Furthermore, the General Transducer performs Named Entity Recognition (NER) on the *Lookup* annotations obtained from the Gazetteer module. The name for i.e. a person can consist of several titles, forenames and a surname, although not all of these are mandatory for the mention of a person. For sequences like "Martin", "Dr. King" or "Dr. Martin Luther King" the Gazetteer module provided the labels title for "Dr.", male forename for "Martin" and "Luther" and as surname for "King" (while "King" was labeled title as well). The NER annotates these sequences with the label *Person* according to its rules (cf. figure 2). Similar name labels such as *Organization* are produced by the NER as well.

The third part of the General Transducer module provides semantic information and assigns semantic roles to the constituents the sentences of the text consist of. The process of assigning semantic roles to constituents is called "Semantic Role Labeling" (SRL) [11]. Sometimes the term "thematic role" is used for "semantic role" [12] so that "Thematic Role Labeling" also denotes that process. The process of SRL links word meanings to sentence meaning by identifying agent-, patient-, instrument-, time-, location-, and other semantic roles. For example, in a sentence of active voice, the subject constituent, the constituent which precedes the verb group, receives the role *Agent*. In case the verb is in passive voice, *Agent* is the constituent following the verb group. The hint for the SRL lies in the word order. In contrast, syntactic and lexical information together often are needed to decide whether there is temporal or spatial information in a sentence. For example, a prepositional phrase that starts with the preposition "in" normally opens a constituent that will be labeled *Location* as in "in a power plant" or a *Point in time* as in "in 9.8 seconds". Gazetteer lists are used to annotate whether a prepositional phrase most probably contains temporal or spatial information. These lists include trigger words, e.g. "minute" or "December" for temporal information, or words like "area" or "space" for spatial information. After classifying constituents as temporal, spatial or suchlike, SRL refines that classification by dividing the temporal constituents into start time constituents (a trigger word would e.g. be "from"), end time constituent ("until") or point in time constituent ("on"). In the same way, SRL sub-classifies spatial constituents as *location* ("at"), *direction* ("towards"), *origin* ("from") or *destination* ("to") and so on.

As the text mining process runs unsupervised, the texts to be processed need to adhere to some requirements which will be explained in the following section.

3 The Wiki-Texts

Within the scope of the EnArgus project, the experts from the energy research domain were assigned to create a Wikipedia-like collection of domain-specific texts in German. The text collection is thought to be processed automatically by the aforementioned information extraction component. The extracted information will subsequently be used by a module building an ontology out of the text collection. The procedures of this module will be described in paragraph IV.

3.1 Restrictions

Not any free text can serve to extent an ontology. In addition to the requirements of information mentioned in the next section, there must be several restrictions to the wording and the syntax used. First of all, the wording should be easy, since the part-of-speech tagger is trained on miscellaneous text corpora and likely fails to assign the correct word category if the wording is too special. The experts therefore are asked to use a common-sense wording which is not too scientific and suitable both for the processing system and the non-expert human users.

Furthermore, there should be as little information as possible enclosed in parentheses or quotation marks. Both of those can potentially occur everywhere in a sentence and interrupt its structure which makes it difficult for the chunker to detect the phrases' boundaries appropriately.

Unambiguous pronominal references are essential for correct information extraction. Pronouns, such as "it" or "this", must clearly refer to the thing they stand for.

3.2 Structure of the texts

The text collection itself will be made public for non-expert users, such as Parliament members, to be used as a glossary. Nevertheless, its priority is to serve as a resource of information about the energy domain for the ontology-building component. Each text describes a specific energy-related item. The texts might more or less be linked to each other. Although this will be helpful for the human user, the ontology-builder does not make use of links.

For an ontology to be built out of a text's content, the text must include information about how the item described needs to be classified in a taxonomy. This means, it must contain statements about the item's respective hyponyms (superclasses) and hypernyms (subclasses). Optionally, the experts can mention individual real-world-objects as examples (individuals) of the abstract item they describe. Furthermore, a text should contain the item's characteristic properties (datatype properties) and relations to other items described (object properties).

4 Ontology Building

In order to build an ontology, two different processes are required which make use of the annotations provided by the information extraction process. First, a preparatory module starts the ontology building and then an ontology builder completes the process by writing an OWL-file.

4.1 Ontology Preparation

Based on the annotations from the information extraction component, the ontology preparation module identifies the structures, such as super- and subclasses, individuals, labels, properties and relations, which will subsequently be used to build the ontology.

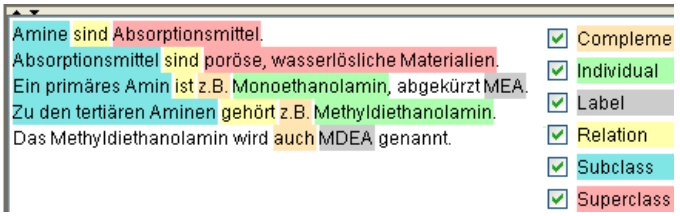


Figure 3. Example text annotated by the ontology preparation module

In a sentence like "*An X is a Y*", the preparation module identifies *Y* as the superior class and *X* as a subclass of *Y*. In a sentence like "*An X is e.g. Y*", *Y* is an individual instead of a superclass and is annotated accordingly. Word sequences like "*referred to as Z*" indicate that *Z* is an alternative expression or abbreviation and must be annotated as a label together with its respective reference. Moreover, the preparation module detects properties of individuals or whole classes. In the phrase "*porous, water-soluble materials*", a class named "material" is identified (the reduced form is obtained by the part-of-speech tagger) as well as the fact that members of this class are able to bear the properties "porous" and "water-soluble". Relations to other objects, such as "part of", are identified by the sentence's verb and annotated *Relation* with meta information about the respective object. Relations between individuals are referred to as object properties, while properties of string-, integer- or boolean values are called datatype properties. Once the preparation is finished, the ontology can be built. This is done by the subsequent module called ontology builder.

4.2 Ontology Builder

The ontology builder requires the previously calculated information from the ontology preparation module. It uses the Protégé's [13] OWL application programming interface [14] to create an OWL-file. The information from class annotations is used to generate the ontology's taxonomy. Superclasses in the taxonomy can bear properties which are inherited to all individuals in all subclasses below that class, i.e. all materials can be water-soluble and porous. Fixed values like datatypes can be defined for these properties, too. Sometimes a new class may be identified by the ontology builder which

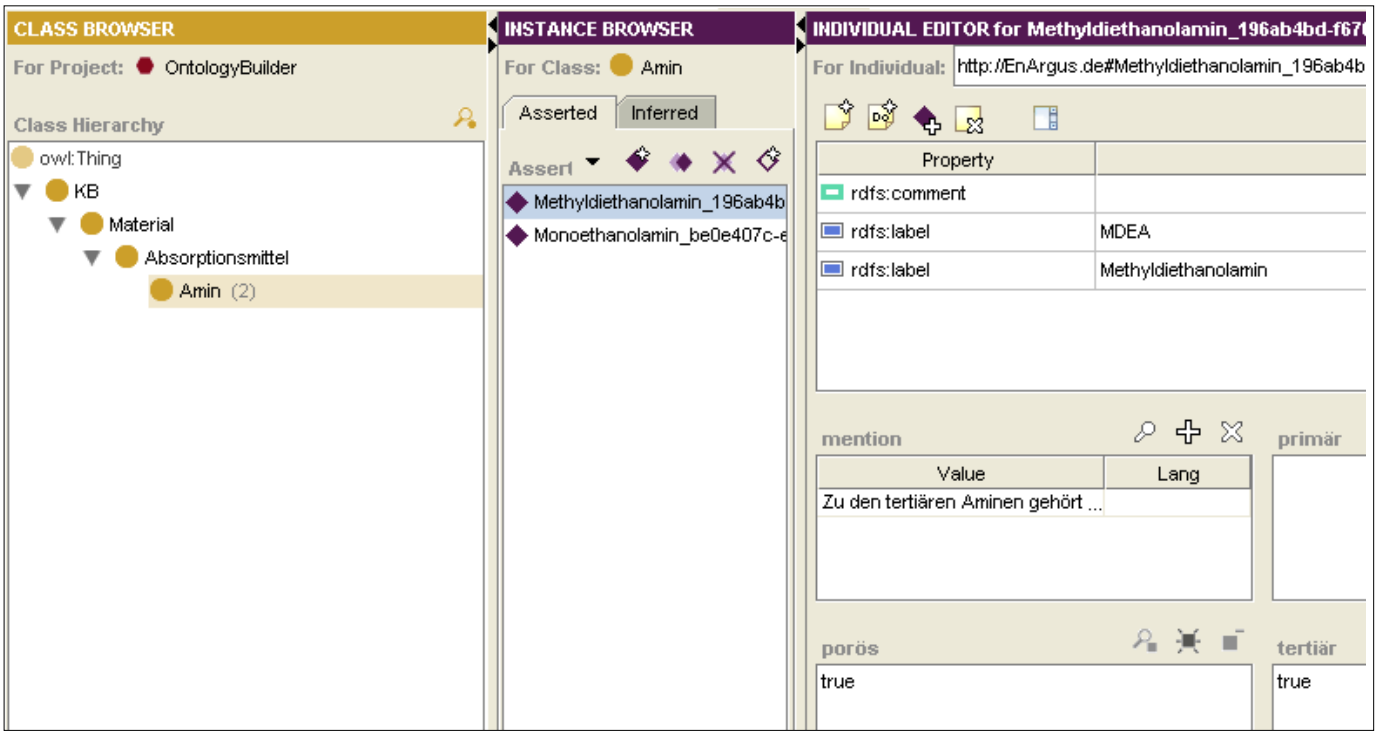


Figure 4. The resulting ontology built out of the example text (fig. 3)

belongs just between two already existing classes. Therefore, the ontology builder is able to paste a new class in the so far built taxonomy. Individuals are attached to their respective classes and each individual inherits all properties from all its superclasses. Properties are also preconfigured with values at class level. This means, e.g. the property "water-soluble" is boolean and for a certain superclass can be set to "true". If there are individuals with the same name but different meanings, a unique ID is assigned to each individual. This enables homonymy and polysemy as well as ontology naming restrictions (names for individuals i.e. may not start with numbers or special characters). Figure 4 shows the resulting ontology from the example text above. On the left (yellow) pane the class hierarchy is displayed. The center (purple) pane indicates the individuals and the right (purple) pane refers to the selected individual's properties and values as well as its labels. The property "mention" is introduced. It displays the context from which the individual, properties, values and labels are derived. With the described text mining component, a domain-specific ontology can be build for the energy research domain.

5 Conclusion

The system presented is a part of an information system for the energy research project EnArgus. An information extraction system with specific wiki texts is presented and it had been explained how to build a specific ontology for this domain. The wiki texts written by energy research experts in a glossary-like format serve as source for that ontology building process. The information extraction system recognizes syntactic and semantic information in the wiki texts. Then, two more processes follow up. The first is a preparation process and the second realizes the proper building. The resulting ready-to-use ontology is part of the EnArgus system that will help to discover all kinds of information about energy research projects. If the user of the EnArgus system is not an expert, she may have not the complete knowledge about all the concepts and technical terms used in reports on energy research. The ontology helps to find appropriate keys for queries. E.g., terms (such as hypernyms or hyponyms) are proposed during search sessions for queries if an entered keyword might be imprecise. Even for experts ontology-based, semantic searching reduces time costs and improves efficiency and quality.

6 References

[1] EnArgus: Central information system for energy research funding. http://www.enargus.de/

[2] GATE: A General Architecture for Text Engineering. http://gate.ac.uk/

[3] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan. GATE: A Framework and Graphical Development Environment for

Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics, 2002.

[4] Mark Hepple. Independence and Commitment: Assumptions for Rapid Training and Execution of Rule-based Part-of-Speech Taggers. Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, pp 278-285. Hong Kong,
2000.

[5] Eric Brill. A simple rule-based part-of-speech tagger. Proceedings of the 1st Applied Natural Language Processing Conference ANLP. Association for Computational Linguistics, Trento/Italy. 1992.

[6] Thorsten Brants. TnT: A Statistical Part-of-Speech Tagger. In Proceedings of the 6th Applied Natural Language Processing Conference ANLP, Seattle, WA. 2000.

[7] Kristina Toutanova, Dan Klein, Christopher Manning, Yoram Singer. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL, pp. 252-259. Edmonton, Canada. 2003.

[8] Helmut Schmid. Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland. 1995.

[9] http://www.aclweb.org/aclwiki/index.php?title=POS_Tagging_%28State_of_the_art
retrieved on Feb 13th, 2012.

[10] Lukas Sikorski, Bastian Haarmann, Ravi Coote. Ontology-driven Information Extraction. Proceedings of the NATO RTO IST-097. Oslo, 2011.

[11] Bastian Haarmann, Lukas Sikorski, Jürgen Ziegler. Applied Text Mining for Military Intelligence Necessities. Proceedings of the 6th Future Security Conference. Berlin, 2011.

[12] Lukas Sikorski, Bastian Haarmann, Ulrich Schade. Computational Linguistics Tools Exploited for Automatic Threat Recognition. Proceedings of the NATO RTO IST-099. Madrid, 2011.

[13] D. L. Rubin, N. F. Noy, M. A. Musen. Protégé: A Tool for Managing and Using Terminology in Radiology Applications. Journal of Digital Imaging, J Digit Imaging. 2007.

[14] Protégé:
http://www.protege.stanford.edu/plugins/owl/api/

Multilevel Approach to Ontology driven Clustering of Web documents

A. Tulika Narang [1], **B.** Prof. R. R. Tewari [2]
[1, 2] Centre of Computer Education, Institute of Professional Studies
University of Allahabad, Allahabad, India

Abstract— *A multilayered knowledge based approach has been proposed to overcome the drawbacks of overload and mismatch. Overload and mismatch are essential issues regarding extraction of information form World Wide Web. The problem of overload occurs when a large number of irrelevant documents may be considered to be relevant. Mismatch occurs when retrieved information is not according to users' expectations.*
It is not easy to obtain right information for a particular user. The layered architecture aims at web content mining. It supports Self Organizing Maps (SOM) clustering of web documents. It is a Neural Network based clustering. The clustering results in groups of relevant Web documents.

Keywords- Web mining, Web content mining, Ontology, Clustering, SOM

1 Introduction

The World Wide Web is a huge, widely distributed, global information repository. But only a small subset of the information is relevant or useful to a user. It is a challenge to find high-quality web pages on a specified topic. It difficult for users to search and retrieve documents that is relevant to their particular needs. Users browse through a large hierarchy of concepts to find the relevant information. The query submitted to a search engine has to wade through irrelevant documents.

The growth of Web has caused of a number of problems with its usage. In particular, the quality of Web search and corresponding interpretation of search results are not according to users' expectations. The retrieved information has drawbacks of mismatch and overload [1, 10].

The focus is to retrieve the most useful and relevant. The challenge has promoted to find methods for effective and efficient searching on the web. The challenge can be overcome by Web content mining. It aims at gathering information from web sources for knowledge discovery.

One solution is to construct meaningful classifications of objects [3 ,4, 5]. The essential application is to group similar objects into classes. Classes or clusters are collections of objects whose intra-class similarity is high and inter-class similarity is low.

Cluster analysis is a technique for multivariate analysis that assigns items to automatically created groups based on a calculation of the degree of association between items and groups [4, 5, 6]. It deals with the organization of a set of objects in a multidimensional space into cohesive groups, called clusters. The groups or clusters which are formed should have a high degree of association between members of the same group and low degree between members of different groups. It is a tool of information discovery. It has the potential to reveal previously undetected relationships between data. The goal of clustering is to create classes or clusters such that the objects within a group are similar and related. The greater the similarity the better and distinct is cluster analysis. In context of web mining cluster analysis aims at grouping Web pages on the basis of document content. It is useful for organizing documents to improve retrieval and support browsing.

Given a collection of web pages D, the notion is to retrieve the best relevant document to particular domain. The documents relevant to a class are grouped together on the basis of relevance function. Each class forms a cluster of similar web documents. The similarity measure is the relevance factor [1, 2]. Relevance denotes how well a retrieved document or set of documents meets the information need of the user. A document belongs to a class only if the relevance value is greater than or equal to a specified threshold.
An approach has been taken that performs clustering of Web documents using ontology [8]. The word "ontology" has a long history in philosophy. It refers to the subject of existence. In the context of knowledge sharing Gruber defined ontology, *a specification of a conceptualization.* That is, ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents. Ontology provides a shared vocabulary, which can be used to model a domain that is, the type of objects, and/or concepts that exist, and their properties and relations. It defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members).

2 Background and Related Work

The vast amount of content on the Internet has made difficult for users to find and utilize information. It is difficult to classify and catalog documents. Traditional web search engines often return hundreds or thousands of results for a

search. There is a requirement for efficient and automated information retrieval methods. The process of extracting useful information from World Wide Web is termed as World Wide Web [10]. It includes the data present in Web pages and data related to Web activity. It is the application of data mining techniques to web-based data for the purpose of learning or extracting knowledge.

Web mining tasks classified as [3, 4]:

- Web content mining
- Web structure mining
- Web usage mining

Web content mining includes the traditional searching of Web pages via content. The goal of web usage mining is to examine web page usage patterns. It uses the Web-log data coming from users' sessions. In this framework, Weblog data provide information about activities performed by a user from the moment the user enters a Web site to the moment the same user leaves it. In Web Usage Mining, the clustering tries to group together a set of users' navigation sessions having similar characteristics

In web structure mining the relationships between web documents is extracted by utilizing the information conveyed by each document's hyperlinks.
The paper focuses on Web content mining. It emphasizes on methods for the automated discovery, retrieval, organization, and management of the vast amount of information and resources available in the Web.
Applying data mining techniques to web page content is referred to as web content mining. It is a sub-area of web mining, partially built upon the established field of information retrieval.
Information Retrieval (IR) is an essential task in web content mining. IR domain is concerned with searching for documents, for information within documents, and for metadata about the documents. It involves the development of sophisticated systems that can act autonomously or semi-autonomously on behalf of a particular user to discover and organize Web-based information. It aims at retrieving the useful and leaving the rest. It is concerned with finding the relevant subset from the available set of documents [10, 11].

Web Content Mining includes many concepts of traditional text mining techniques. Clustering is one such technique. To enhance the retrieval process the basic text mining is augmented with various data mining tasks such as clustering and classification.

Clustering groups similar documents together to make information retrieval more effective. Web document clustering methods identify inherent groupings of pages so that a set of clusters is produced in which clusters contain relevant pages to a specific topic. The irrelevant pages are outliers and are not included in any group. The clustering methods group the documents into clusters. Each cluster or group represents some topic that is different than those topic represented by the other groups. It is helpful for discrimination, summarization, organization, and navigation for unstructured Web pages. It aims at improving the traditional informational retrieval performed by search engines [12].

The objective of web content mining can be achieved by [4, 6]:

- Finding the schema of Web documents
- Building a web warehouse
- Building a knowledge base
- Building a virtual database

The objective of web content mining is to retrieve the most useful and relevant as compared to the traditional Information Retrieval systems. It is achieved by building a multilayered knowledge base. The knowledge base supports clustering of Web documents. Clustering is the process of collecting Web sources into groups so that similar objects are in the same group and dissimilar objects are in different groups. Clustering on the Web has been proposed based on the idea of identifying homogeneous groups of web documents. A neural network approach to clustering is proposed [3, 5]. It represents each cluster as an exemplar. An exemplar acts as a prototype of the cluster. New objects can be distributed to the cluster whose exemplar is the most similar. Similarity can be computed using a similarity function.

A self-organizing map is a neural network approach to unsupervised clustering. It has only one layer, the output layer. In this layer, neurons are organized according to a topology where a neuron may have 4, 6, 8, or more neighbors.

It uses competitive unsupervised learning. In competitive learning, nodes are allowed to compete and the winner takes all. Learning is based on the concept that the behavior of a node should impact only those nodes and arcs near it. Weights are initially assigned randomly and adjusted during the learning process to produce better results. During the learning process, hidden features in the data are uncovered and weights are adjusted accordingly.

3 Ontology driven clustering of web documents

3.1 Multilayered Knowledge Base

The objective of the paper is achieved by defining a multidimensional view of the World Wide Web. To provide a

multidimensional hierarchical view of the Web a multilayered web information base is proposed [3].

A multilayered knowledge base is represented as in Fig1.

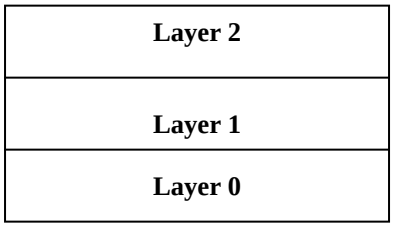


FIGURE 1: *Mulilayered Knowledge Base*

- Layer 0 represents the Web itself.
- Layer 1 is the Web page descriptor layer. It contains descriptive information for pages on the Web.It contains information related to Web documents such as address of the page, keywords, timestamp; access frequency. The attributes are useful in web mining.
- Layer 2 represents the clustering process. The SOM clustering algorithm clusters web documents relevant to a particular domain. It performs clustering by several units competing for the current object. The unit whose weight vector is closet to the current object becomes the active unit. It is a neural network approach to clustering and is closely related to processing done by human brain.

3.2 Ontology For Document Representation

An ontology describing the domain is built using either text corpora or manual definitions or both. Then documents are indexed. The documents are numerically represented by vectors whose dimensions correspond to indexing units. The vectors store the weight of the indexing unit and are input to the clustering algorithm.

Ontology construction results in a hierarchical structure comprising of concepts and the relationships between them. This structure includes the inter-relationships between concepts.
An ontology can be represented as a tuple , **T** [3,7].

$$T= (L, F, C, H, Root)$$

Where,

L: Lexicon that contains set of terms

C: Set of concepts

F: Reference function that links a set of terms to the set of concepts they refer to

H: hierarchy that contains relationship between concepts

ROOT: Top concept and belongs to set C

The following steps are performed to generate vector of documents from ontology structure:

- Identify and extract terms from documents. This represents the domain lexicon.
- The lexicon is mapped to a ontology structure.
- When ontology is available documents are indexed.
- Indexed documents are used to generate vectors of documents that are input to the clustering method.

The sequence of steps can be represented as in Fig 2.

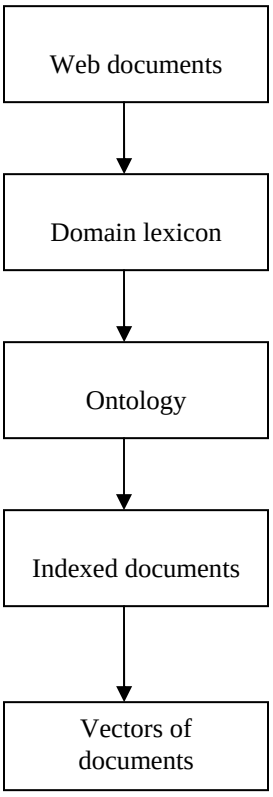


FIGURE 2 Steps to generate vector of documents

3.3 Clustering Web Documents

Clustering is a data analysis technique. It deals with the organization of a set of objects in a multidimensional space into cohesive groups, called clusters. It is similar to classification. But unlike classification groups or clusters are not predefined [4]. In classification data items are assigned to a predefined category based on a model that is created from pre-classified training data (supervised learning).

The goal of clustering is to separate a given group of data items (the data set) into groups called clusters. It emphasizes that items in the same cluster are similar to each other and dissimilar to the items in other cluster. In clustering methods no labeled examples are provided in advance for training. It is also called unsupervised learning).

The Web consists of a variety of Web sources. In order to facilitate data availability and accessing, and to meet user preferences, the Web sources are clustered with respect to a certain parameter or characteristic such as their popularity, structure, or content. Clustering on the Web can be one of the following types:

• Web User Clustering: The establishment of groups of users exhibiting similar browsing patterns. Such knowledge is especially useful for inferring user statistics in order to perform various actions such as market segmentation in e-commerce applications, and personalized Web content for users. This type of clustering helps in better understanding the users' navigation behavior.

• Web Document Clustering: The grouping of documents with related content. This information is useful in various applications, for example, in Web search engines
This improves the information retrieval process (i.e., clustering Web queries). In addition, the clustering of Web documents increases Web information accessibility and improves content delivery on the Web.

A clustering method relies on four concepts:

- model of data to be clustered,
- similarity measure,
- cluster model, and
- clustering algorithm that builds the clusters using the data model and the similarity measure.

The proposed approach is based on the vector space model. The similarity measure is the relevance factor. The clustering algorithm applied is SOM based clustering based on neural network approach of clustering.

To represent text and web document content for clustering a vector-space model is used. The process of clustering documents begins with selecting the type of the characteristics or attributes (e.g. words, phrases or links) of the documents on which the clustering will be based and their representation. Clustering is then performed using as input the vectors that represent the documents.

In vector space model a document is represented as a vector of the terms that appear in all the document set. Each term in a document becomes a feature dimension. Each feature vector contains term weights of the terms appearing in that document. The term weighting scheme is usually based on tf . idf method in IR [3].

The value assigned to each dimension of a document may indicate the number of times the corresponding term appears on it. It is a weight that takes into account other frequency information, such as the number of documents upon which the terms appear. This model is simple and allows the use of traditional machine learning methods that deal with numerical feature vectors in a Euclidean feature space.

The problem is that traditional data mining methods are often restricted to working on purely numeric feature vectors due the need to compute distances between data items or to calculate some representative of a cluster of items (i.e. a centroid or center of a cluster), both of which are easily accomplished in a Euclidean space. Thus either the original data needs to be converted to a vector of numeric values by discarding possibly useful structural information (which is what we are doing when using the vector model to represent documents) or we need to develop new, customized methodologies for the specific representation

The paper presents a SOM based clustering technique [3]. It is a neural network clustering method.

Clustering of Web documents precede with preprocessing textual documents. Preprocessing involves two essential tasks:

1. Representation of the domain
2. Calculating the importance of semantic element.

The domain is represented using vector space model. Each document is represented as a feature vector, whose length is equal to the number of unique document attributes in the collection. Each component on that vector has a weight indicating the importance of each attribute in the characterization of the document. Usually, these attributes are terms that are extracted from the document using IR techniques.

A vector representing the neuron is called neural vector. This vector has same number of dimensions as the input vectors.

These vectors are inputs to the clustering process. To generate vectors from documents an indexer is required. Indexing is dependent on term extraction. A term represents a semantic element having meaning in the particular domain. A term can be a keyword in the document. The importance of a term is measured using term frequency (TF) or Term Frequency-Inverse Document frequency (TF.IDF) [1,3]. TF is the number of occurrences of the term in the document. The greater the value of TF the more important is the term. TF.IDF is an improvement over TF.

The importance of a semantic element is based on its frequency of its occurrence in the document. A semantic element i .e. a term is important to a document only if it is present in the document. A weight value 0 associated with a term implies that the term is not at all important to the particular document. In other words, a document is relevant to a particular domain only if it contains important semantic elements.

The significance is computed using a relevance function.

Relevance function R(d) for web documents is defined as [1,2] :

$$\mathbf{R(d) = \sum pr_{\beta}(t)\ \rho(t, d)} \qquad (1)$$

Where $\rho(t, d) =$ **1 if t ε d , otherwise 0**

The relevance function is based on the following two dimensions:

- Exhausitivity: It describes the extent to which pattern or topic discusses what users want.
- Specificity: It describes the extent to which the pattern or topic focuses on what users want.

A pattern (P) is a set of term frequency pairs. Each term or keyword is an individual semantic element. A measure **support(P)** is used to describe the extent to which the pattern is discussed in the set of documents under consideration [9]. The greater the support is, the more important the pattern is.

$\mathbf{pr_{\beta}(t)}$ is directly propositional to **support(P).**

β is a mapping that explicitly describes the relationship between patterns and the common hypothesis space.

4 Conclusion and Future Work

The paper represents a multilayered approach to web content mining. It is an attempt to overcome the limitations of web content mining. The proposed solution clusters set of web documents into classes or groups. Clustering web documents separates unrelated pages and clusters related pages (to a specific topic) into semantically meaningful groups. It is useful for summarization, organization and navigation of unstructured Web pages.
Each cluster is semantically similar on the basis of terms. A relevance function computes similarity measure. A document is relevant to a particular domain if it contains terms or keywords essential in describing the domain.

The formal implementation of the proposed algorithm and its performance is in process. We will implement the approach for designing a domain specific web search engine. The search engine will use a web crawler to index only web pages that are relevant to a predefined topic or set of topics. The search engine will search pages that are relevant to a particular domain. It will contrast general search engines that index large portions of World Wide Web.

5 References

[1]Yuefeng Li and Ning Zhong, "Mining Ontology for Automatically Acquiring Web User Information needs", IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 4, April 2006

[2]Xiaohui tao, Yuefeng Li and Ning Zhong, "A Personalized Ontology Model for Web Information Gathering", IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 4, April 2011

[3]Hector Oscar Nigro, Sandra Gonzalez Cisaro, and Daniel Hugo Xodo, Data Mining with Ontologies-Implementations, Findings, and Frameworks, IGI Global, ISBN 978-1-59904-618-1

[4]Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques,ISBN 81-8147-049-4

[5]Margaret H. Dunham and S. Sridhar, Data Mining–Introductory and Advanced Topics,Second Impression, 2007, ISBN 81-7758-785-4

[6]Richard J. Roiger and Michael W. Geatz, Data Mining-A Tutotrial-based Primer,First Indian Reprint, 2005, ISBN 81-297-1089-7.

[7] Wanlong LI, Dayou Liu, Shanhong Zheng, Suyun Jiao," A Novel Computational Approach to Concept Semantic Similarity", International Conference on Computer, Mechatronics, Control and Electronic Engineering 2010.

[8] Ernesto William De Luca, Andreas Nürnberger, "Using Clustering Methods to Improve Ontology-Based Query Term Disambiguation" International Journal Of Intelligent Systems, Vol. 21, 2006

[9] Yuefeng Li and Ning Zhong," Capturing Evolving patterns for Ontology based Web mining", International Conference on Web Intelligence, 2004.

[10] Y. Li and N.Zhong,"Web Mining Model and Its Application on Information gathering," Knowledge Based Systems, vol 7, 2004

[11]R. Baeza Yates and B. Ribeiro Neto, Modern Information retrieval, Addison Wesley, 1999

[12] B. Liu and K. Chang, SIGKDD Explorations, special issue on Web content mining, vol 6, no. 2, 2004

Turkish Query Engine on Library Ontology

Sadi Evren SEKER
Department of Computer Engineering, Istanbul University, Istanbul, Turkey
academic@sadievrenseker.com

Abstract *Purpose of this project is implementing conversational software to interface dialog based sentences between the user and a library database.*

This software implemented with a special expertise on library dialogs. The number of possible library dialog sentences is limited and this project covers almost all of these possible sentences. The input sentences are accepted as Turkish and a flexible management system added for further additions. For example any sentence missed on this project can be added with a simple entry on the YACC file.

The technology utilized during this project is YACC and LEX implementation on LINUX. Also the database of the project is implemented over MySQL. LEX and YACC produces C source codes and the functionality of semantic processing and the database queries are also implemented in C language.

One of the hardest part of this project is implementing Turkish language capability over C programming environment on LINUX. All the technological modules of this project which are MySQL, C, LEX, YACC and ZEMBEREK created different problems with the Turkish inputs. During these problems I have searched Internet for the Turkish input implementations of LEX and YACC or the MySQL connection through C and as a result of my findings this project is the first time implementation of Turkish characters sets by LEX, YACC and MySQL at the same time on LINUX.

One of the most important achievements after accomplishing this study is the flexibility of the input sentences. Anybody can add a new grammar rule to the YACC file buy obeying the regular expression structure of YACC. After a successful addition the project will search for this new addition in the input sentences and the answers related to this input will be produced.

Keywords: Natural Language Processing, Ontology, Morphological Analysis, Syntactic Analysis.

1 Introduction

Aim of this project is implementing conversational library software, which will be capable of understanding the user inputs and reply back the user with the desired information. The user inputs can be anything with no limitations of course, but the software is only responsible to understand the library based dialog inputs.

The only data source of this project is a database, implemented to keep the information of writers and books. The database implementation kept as simple as possible to make the conversational part of project more challenging.

The project implemented in a flexible manner, so any sentence pattern can be added to the conversational module.

Almost all for the bots are developed for English and the biggest difference of this project and the current bots are the language they have developed. Also they have mostly frozen sentence structures instead of a formal approach with morphological, syntactic and semantic levels. They mostly try to find out a NLP parsing by using the prepared sentence templates and when a user input matches this template the related answer is displayed. This is the most important second part of this project from the projects above.

Also the database of the above bots is very limited and they do not have a chance to measure their successes. Since the world has unlimited number of natural language sentences (theoretically) the above bots can never be measured in a way of all possible inputs.

Contrarily this project has a limited number of input possibilities and the maximum possible sentences can be touched one day. Of course during this study the all possible sentence alternatives are not outlined but this is a possible case for the future studies.

In the case of finding the possible alternatives the performance of success will be able to measure by using the current implementation.

Of course the measurement is an essential element in engineering studies so I have tried to find out a way of implementing these possibilities. The solution is dividing the possible input sentences into two groups. In the first group the sentences will be used for training and the code will be developed. The second group of input sentences will be kept for the testing purposes and the performance of success will be measured by using the outputs of these sentences.

2 Design of Project

All the possible inputs will be covered and a BNF [1] representation will be created in this chapter. Also a finite state machine will be drawn depending on the BNF representation of the input sentences.

The design phase covers the details of technological selections and their adaptations to the project.

2.1 Layered Approach and Technological Background

The project can be separated into 4 layers as demonstrated below:

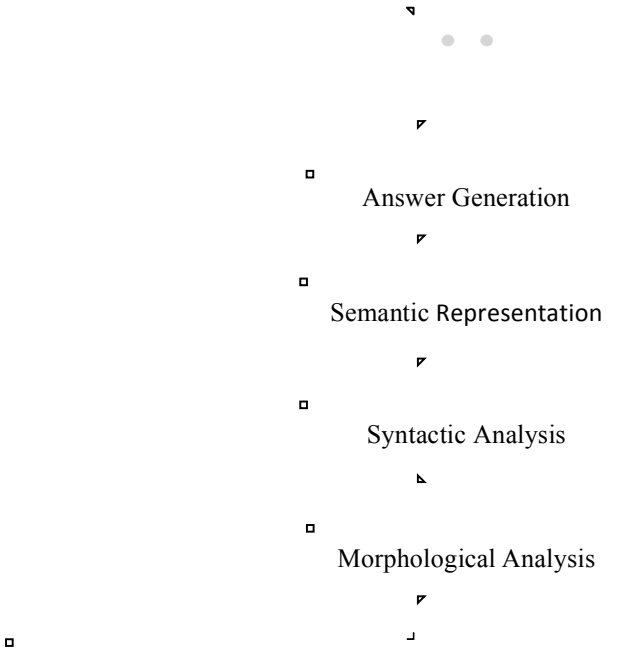


Fig. 1. Layered representation of the Project

Above figure represents the three well-known layers. The natural language processing starts by the morphological analysis the order of words is parsed in the syntactic phase and finally a semantic representation is found out. The semantic representation is of course not a user-friendly representation. The aim of this representation is only keeping track of the meaning of the input sentence. The project gets this semantic representation and tries to find out a possible answer for the user query and generates an answer in natural language again.

2.2 Morphological Analysis

The morphological analysis [2] is the phase of getting the meaning of the words. The word meanings are mostly depends on the root of the word. Of course the suffixes may shift the meaning of the word. During the studies the project has got 3 phases.

1. The pre yacc phase
2. Yacc integration phase
3. Morphological Regular Expressions phase

In the pre yacc phase the integration of lex and zemeberek has taken place and the words have been separated into three categories:

- Frozen words
- Zemberek processed words
- Database queried words

The frozen words was the hard coded words which does not has any suffixes and does not match in the database.

Zemberek processed words was accepted words in the syntax of the project but they do require some morphological analysis because of many possibilities with the words.

For example consider the below sentences:
Beyaz isminde kitabı yazan kimdir
(Who has written the book named "Beyaz")
Beyaz isminde kitabı olanlar kimlerdir
(Whose book has a name of "Beyaz")
Beyaz isimli kitabı olan kimdir
(Who has a book named "Beyaz")
Beyaz ismindeki kitabı yazan kimdir
(Who is the writer of the book "Beyaz")
Beyaz ismindeki kitabı yazanlar kimlerdir
(Who are the writers of the book "Beyaz")
Beyaz isimli kitapları olanlar kimlerdir
(Who are the writers with the book "Beyaz")

All the above sentences have almost same syntactic order with same semantic meaning. So the user entering any of the above sentences means the same thing for each of the above options.

The morphological meaning gets a great variety by the above sample for example all the below words have the same semantic:

Kitap*(book)*, kitabı*(the book)*, kitaplar*(books)*, eseri *(the work)*, eserleri *(works)*, eser *(work)*, romanı *(the novel)*, romanları *(novels)*, roman*(novel)* … etc.

Similarly from the above sentences also the below words has the same semantic representation:

Isminde *(in the name)*, ismindeki *(the named)*, isimli *(with the name)*

So the morphological analysis should take care of the word varieties and return the same semantic representation for the words in the same meaning.

Finally the third word type in the morphological analysis is the database queried words. These are the free words for any possible variety. So the user can use almost any word in this part give the name of a writer or a book.

The pre-YACC phase was the querying phase from the database. So the user enters a word in the appropriate place of the sentence the word gets queried in the database.

The database returns the word result and morphological analysis gets sure that the word is a writer or book phrase.

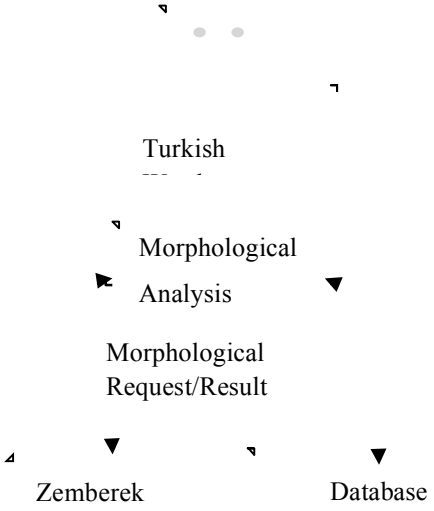


Fig. 2. Deployment of morphological analyzer

Above figure demonstrates the deployment of morphological analyzer between user, zemberek [5] and database in the pre yacc development phase.

After the failure of the yacc integration the above deployment has changed.

2.3 YACC Integration Phase

After integrating YACC [3] into the project, the queries have been carried to the YACC to generate an answer to the inserted question. The questions are parsed into the words and the LEX was only processing these words and producing the frozen results.

During this phase the zemberek integration and the database connection was completely replaced with the frozen words.

Fig. 3. YACC integration development phase of morphology

Above figure demonstrates the processing of the user input in the LEX[3]. Of course the result of the above process was carried out to the syntactic analysis which was YACC in this step.

The problem occurred in the possibilities of the words. In the previous section the list of all possible syntax entries with great variety requirement on the morphology has forced a solution to the morphological integration.

The problem was the number of frozen words entered in the LEX.

This problem has solved in the final version with morphological regular expressions phase.

2.4 Regular Expressions in Morphology

This final version of morphology solution can be understood better from the below example:

Let's consider the below frozen word entries in the LEX file from the previous section:

- Eseri (His art work)
- Eserini (The art work of him)
- Eseriyle (With the art work of him)
- Eserleri (his art works)
- Eserler (art works)

Before this development phase all the above words was accepted by separate entries in the LEX file. After the regular expression improvement the whole above sentences was accepted by a single rule:

```
"eser"[a-zğüıöçş]* {
      return KP;
}
```

Also this reduces the number of syntactic rules since each variety above is represented by only a single entry of KP return value.

Furthermore this improvement gives another possibility of reducing syntax by removing the below word alternatives:

```
"roman"[a-zğüıöçş]* {
return KP;
}
"hikaye"[a-zğüıöçş]* {
return KP;
}
"kitab"[a-zğüıöçş]* {
return KP;
}
```

The syntax would have been covered all the possible entries of above word roots as a rule.

Besides the above improvement a major obstacle has faced during this implementation. The shorter roots were getting the priority which causes problem.

For example consider the word root "yaz" and "yazar" in the same LEX file entry. The "yaz" root has the priority since it is shorter than the "yazar" entry.

This problem causes a masking of "yazar" entry in the LEX file.

The solution is quite simple. The order of the LEX file entries defines a priority on the rules. For example the "yaz" root should be kept below the "yazar" root so the "yazar" root will get priority and only if the "yazar" root

has not matched than the LEX goes beyond and checks for whether the "yaz" root matches for the current word.

2.5 Syntactic Analysis

This chapter covers the syntactic analysis of the project. The syntactic analysis is responsible from the order of the words. The word order is parsed in this phase by depending on the morphological semantic and the meaning of the words.

For example the word meanings can vary from order to order, so the syntactic analysis goes through these word order alternatives and makes a comparison between the expected word order and the input word order. In the case of a match with the expected and input word order, the syntactic analysis generates a semantic representation for the answer generation phase. In fact this last phase can be integrated by the syntactic analysis since the possible outputs are already separated at this level.

The design of syntactic analysis can be separated into two major parts. The first part was mostly targeting the ontological queries, so the failure of the LEX integration.

- Ontological Queries phase
- Database Query phase

2.6 Ontological Queries Phase

The first phase gets queries the ontological questions. Unfortunately the implementation of ontological queries[4] was not possible because of the time limitations. The possible ontology queries were quoted below:

Below list holds the sample inputs targeted to be understood by the software. (Library Ontology)

Sample Sentences and their syntactic frames and semantic representations:

Before formulating the possible sentences in library dialog the formal definition of word phrases are given below:

Book Phrases (BP) holds the information about the name of the book. The below list holds the possible variations of the book names in Turkish sentences.

```
BP > BN
BP > BN Roman [ı|nı|lar|ları|larını]
BP > BN Kitab [ı|nı|lar|ları|larını]
BP > BN Eser [ı|nı|lar|ları|larını]
BP > BIR BP
BP > HERHANGI BIR BP
```

Writer Phrases (WP) holds the information about the name of the writer. The below list holds the possible variations of the writer names in Turkish sentences:

```
WP > WN
WP > WN Yaz [dığı]
WP > BIR WP
WP > HERHANGI BIR WP
```

Also library dialogs may contain the translation of the books. For example a book may be translated into many languages and the customer may request any copy of these languages. Below phrases are the samples of the translation phrases which may be contained in a possible library dialog in Turkish:

```
TP> TN [si]
```

Questions also holds several question words like who, where, when, etc. Below phrase is responsible to detect such phrases:

```
QP> [kaç [a|da|ı|ncı] |ne|hangi [si]
|var mı|yok mu
```

Other general purpose phrases can be listed as below:

```
VP -> Verb Phrases (consists of a verb
and possibly several adverbs)
NP -> Noun Phrases (consists of a noun
and possibly adjectives)
```

According to above word phrases the possible formulation of some target sentences are given below:

```
LP>BP TP NP QP VP?
```

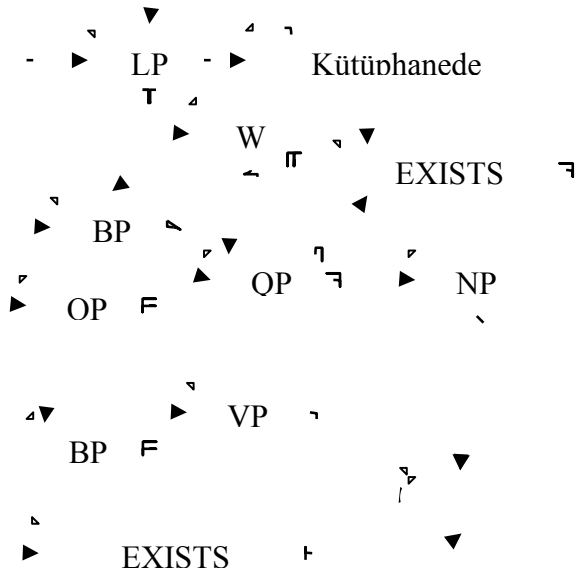


Fig. 4. FSM of library phrases

Figure 4 holds the finite state machine of the library phrases. Figure 4 is drawn according to the above samples.

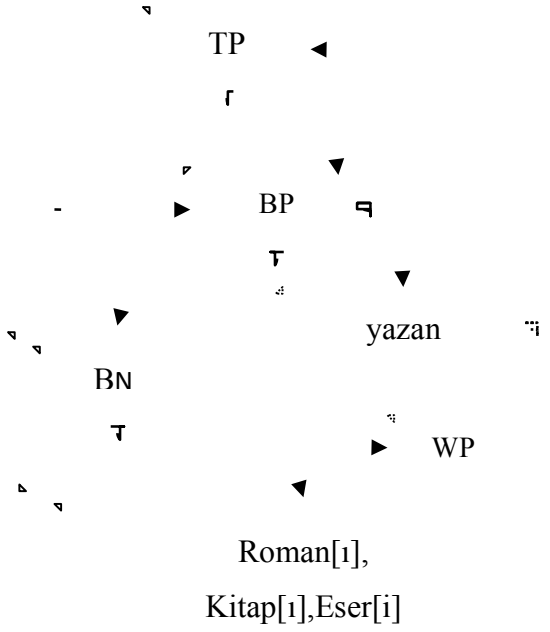


Fig. 5. FSM of Book Phrases

Figure 5 holds the book phrases and the finite state machine is drawn according to the above samples.

A book phrase can be translated to the writer phrase by the keyword "yazan"*(writer)*. In this case the database search is executed to translate the book phrase to the writer phrase.

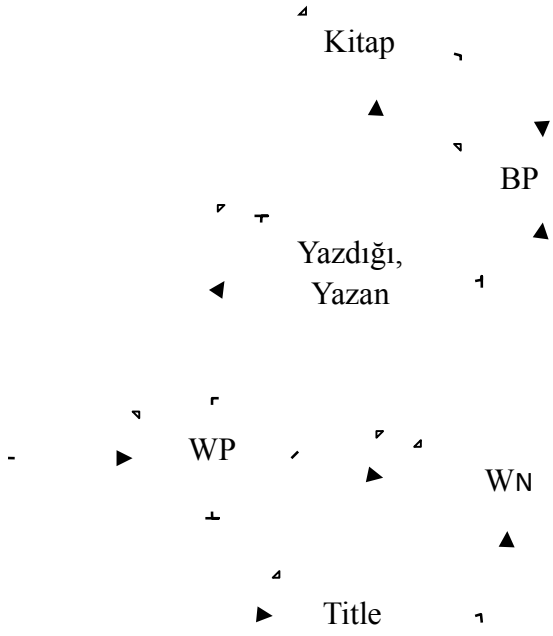


Fig. 6. FSM of Writer Phrases

Above figure holds the writer phrases and the possible variations of the writer phrases. A writer phrase can be translated into the Book Phrase by adding several words besides the writer phrase.

Also a title can be appended before writer noun as a defining statement.

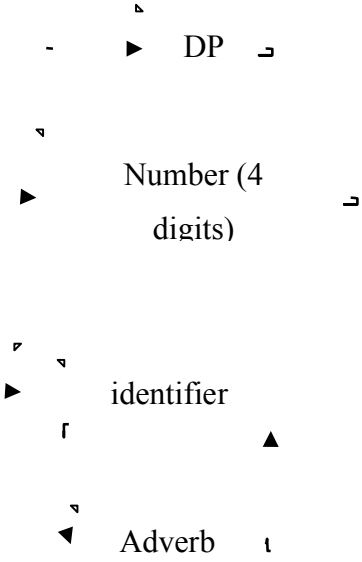


Fig. 7. FSM of Date Phrases

Above figure holds the date phrases of the possible alternatives. A date phrase should contain a number with 4 digits. Also an identifier or an alternative adverb can be added after the date phrases. For example in a date phrase like "1976 yılında yayınlanan" *(in the year 1976)* the number 1976 is considered as a number and "yılında" is considered as an identifier. The word "yayınlanan" is an adverb identifying the date.

A question phrase is a frozen list of words. The list consists of the above grammar, which is also quoted below:

```
QP> [kaç [a|da|ı|ncı] |nere [de|ye]
|nasıl|ne zaman|ne kadar|ne|hangi [si]]
```

Also a translation phrase is again a list of frozen words. Such as "ingilizcesi, almancası, rusçası".

Verb phrases and the noun phrases are consist of the morphological analysis.

2.7 Database querying phase

In this phase of maturity the database queries are executed in the YACC file.

The query details and the implementation of functions will be explained in the implementation chapter. This sub section covers the details of the Rules of the YACC file with respect to the database querying phase.

Above rules are designed for the database querying phase. Each of the above finite state machines represents a word group in the syntactic analysis.

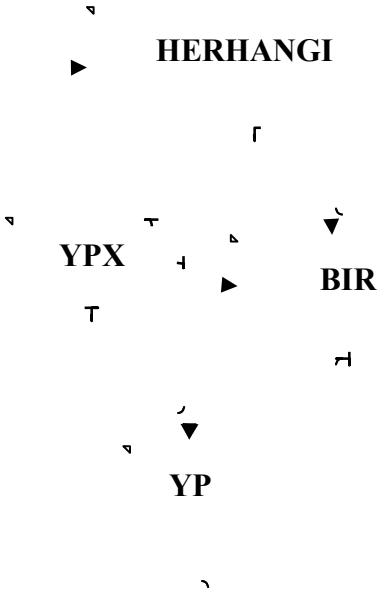


Fig. 8. Finite State machine of writer phrases

Above figure holds the possibilities of the writer phrase. The writer can be a writer name (YP in above figure) or can take any of the words “Herhangi” or “Bir”.

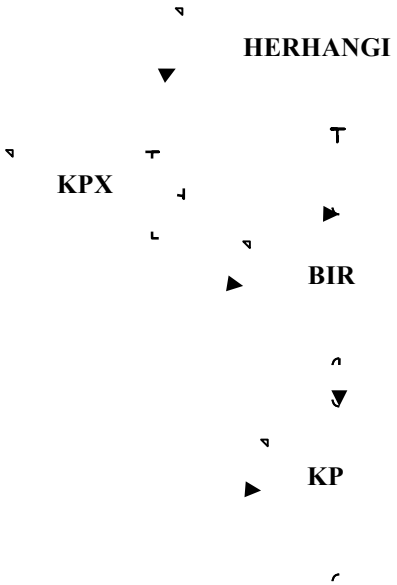


Fig. 9. Finite State machine of book phrases

Above figure is similar to the writer phrase the only difference is the replacement of the writer with the book.

The book name can again get any of the words “Herhangi” or “Bir”.

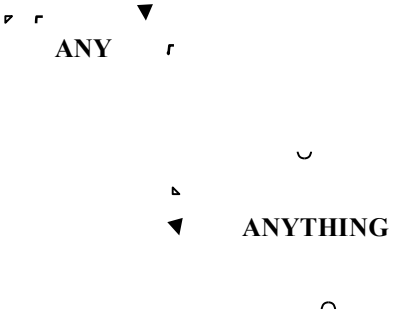


Fig. 10. FSM of Writer and book names

Writer and book names should end by ANYTHING which is a lex entry in the LEX file. And also the lex entry can get any number of pre words.

```
Orhan
Orhan Veli
Orhan Veli Kanık
```

For example all the above writer entries should be fetched by the above rule. So the above rule is implemented flexible to the number of words.

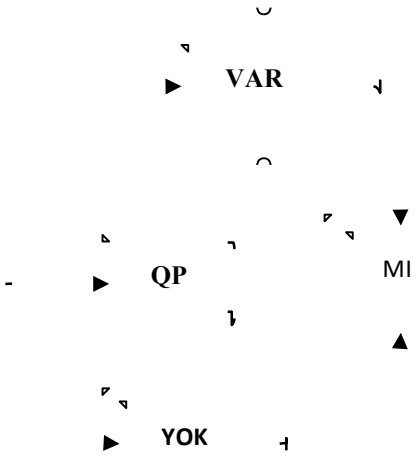


Fig. 11. FSM of Question phrase

Above finite state machine holds the accepted question phrases of existence. The user can query whether something exists or not by the above FSM.

The final view of the accepted phrases is given below:

2.8 Database Design

This chapter covers the details of the database design and implementation. MySQL [6] is selected for the database management system because of its flexible and easy to maintanance behaviour. Also this project is database technology independent, so any other technology can be easily substituted with the MySQL technology.

Below is, the database scheme of the required tables and the columns of each table.

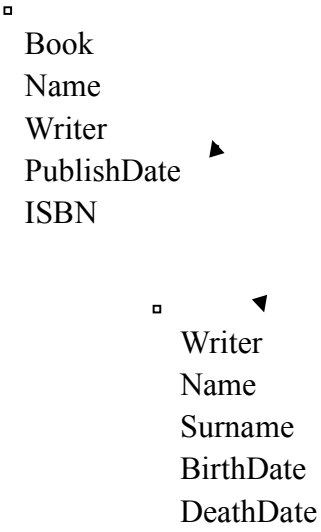


Fig. 12. Database scheme of the book automation.

It is obvious that the above scheme is a simple databas structure with two tables. The database structure keeps the relation between writer and the books. Also the cardinality of the relation between books and the writers does not provide a many-to-many relation. The relation is a simple one-to-many relation. For example a book can be written by only 1 writer. In the case of a book with multiple writers, the system does not handle such cases.

3 Conclusion

During this project, ZEMBEREK has implemented using JAVA technology on both LINUX and Windows environments and because of the problems on connection between YACC, LEX and Zemberek this technology is solved by implementing regular expressions over LEX file.

The LEX file input parser and regular expressions are implemented and modified for the Turkish character set and the connection between LEX and YACC is solved by carrying out these Turkish input characters.

Another technological achievement during this project is implementation of MySQL over the LINUX with Turkish input. Also creating connection to the MySQL via C programming was another problem and solved by installing the MySQL source packages on the LINUX.

Finally the project is capable of running the LEX, YACC, MySQL and C codes compiled together with Turkish input support.

After these technological installation problems the implementation problems has solved by writing several query functions. Also another achievement is reached by adding several input sentence patterns to the YACC file. YACC file keeps the project flexible because of its input pattern support. The LEX file keeps the input flexible by its regular expression support and by adding the flexibility of the LEX to the YACC the project can almost cover any input sentence in Turkish. The only problem of such flexible environment is preventing one of the rules already inserted by adding another rule to the LEX or YACC files.

Another problem related to the flexibility was the database records. The book names on database can keep almost any of the inputs including the frozen words in a YACC input pattern. The problem is solved by adding a flexible recursive YACC rule.

Besides the above technological and designing achievements this project was also an opportunity for me to implement my theoretical studies during undergraduate to a real life project.

For example management of this project is done parallel to the linear model in Software engineering project. The programming skills from introduction to programming and data structure courses helped me to implement the string manipulations' and the data structures implemented during the functions of YACC.

Also the installation and management of modules over LINUX operating system has been achieved in operating systems courses. Rss3@nz

ppdif

Besides the technological and design achievements and the opportunity to practice the above courses I have implemented a relatively large scale project first time. I hope this experience will help me in the future engineering cases.

ACKNOWLEDGMENT

This project has been supported by the research department of Istanbul University, project number YADOP-16728.

4 References

[1] On the Theoretical Foundation of Meta-Modelling in Graphically Extended BNF and First Order Logic Hong Zhu Theoretical Aspects of Software Engineering (TASE), 2010 4th IEEE International Symposium on Publication Year: 2010 , Page(s): 95 - 104

[2] Arabic Morphological Analysis: a New ApproachSonbol, R.; Ghneim, N.; Desouki, M.S. Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on Digital Object Identifier: 10.1109/ICTTA.2008.4530014 Publicat ion Year: 2008 , Page(s): 1 - 6

[3] Simple calculator compiler using Lex and YACC Upadhyaya, M. Electronics Computer Technology (ICECT), 2011 3rd International Conference on Volume: 6 Digital Object Identifier: 10.1109/ICECTECH.2011.5942077 Pu blication Year: 2011 , Page(s): 182 - 187

[4] Ontological queries: Rewriting and optimization Gottlob, G.; Orsi, G.; Pieris, A. Data Engineering (ICDE), 2011 IEEE 27th International Conference on Digital Object Identifier: 10.1109/ICDE.2011.5767965 Publicati on Year: 2011 , Page(s): 2 - 13

[5] Information retrieval from turkish radiology reports without medical knowledge Kerem Hadımlı, Meltem Turhan Yöndem FQAS'11: Proceedings of the 9th international conference on Flexible Query Answering Systems, October 2011

[6] MySQL: lessons learned on a digital library Di Giacomo, M. Software, IEEE Volume: 22 , Issue: 3 Digital Object Identifier: 10.1109/MS.2005.71 Publication Year: 2005 , Page(s): 10 - 13

Answering Spatio Temporal Multimedia Queries Using Configurable Ontology

Sunil Kumar Kopparapu, Arun Pande
TCS Innovation Labs - Mumbai, Yantra Park, Thane (West), Maharastra, INDIA
SunilKumar.Kopparapu@TCS.Com

Keywords: spatio-temporal multimedia, question-answering, ontology

Abstract: With the mobile phone becoming ubiquitous it has become easy to find the channel that can assist people in rural areas query for answers from experts in the urban areas. However, the number of experts in any domain are significantly outnumbered by the people seeking information in that domain; this necessitates a question answering (QA) system. In this paper, we discuss a system that is required to answer a query that has a spatial, a temporal and a multimedia component. We concentrate on the agricultural domain due to our previous experience in building a platform (mKRISHI - An Agro Advisory Platform) that enables rural farmers seek expert answers. The focus of the paper is in (a) formulating the multimedia query answering problem and (b) proposing a configurable ontology to assist answering a query.

1 INTRODUCTION

Question answering (QA) systems (Lopez et al., 2007) have been around for a while now and the intent of building these systems is to enable a machine automatically answer a query, generally posed in natural language (Wu et al., 2011). Most of the research work is concentrated on how to derive the intent of the query and then to ensemble and construct an answer from a set of answer paragraphs aided by a domain knowledge base or an ontology.

In recent times, especially with the proliferation of mobile phones it has become easy for folks in the rural regions seek answers from experts in the urban regions on a wide range of topics. From the Indian context, agricultural related information is in high demand where folks in the rural areas expects answers from the experts who are in the urban areas. While a human expert can answer a query without much problem, the relevance of the answer increases with the knowledge of current and past environmental conditions from which the query originated. To enable experts have access to environmental information in addition to the query, we built a platform called mKRISHI, a Mobile Agro Advisory Platform (TCS, 2012), which an expert a view of the environment along with the query so that the expert can answer the query very precisely. The query from the rural folks generally consists of audio, video, photograph in addition to the textual query. In several domains (more so in agriculture) the answers to the same query is different for different spatio-temporal[1] conditions and could be very personalized to the person seeking the answer. In brief, same or similar queries coming from different geographic (spatial) regions could evoke different answers. For example, a query about *Should I apply pesticide now?* coming from a farmer in a region that is expecting mild showers in the next few days need to be addressed differently from the same query coming from a region that has no possibility of rain in the next couple of weeks. Additionally, same query coming from a farmer who own one acre of land and a farmer who own hundred acres of land needs to answered differently. In such a scenario, very often the answers that an experienced expert would provide would be based on several parameters (spatio-temporal) including specific personal details of the person seeking the answer.

Like in any scenario, the number of human experts are considerably outnumbered by the people seeking information, this necessitates the need for an automatic question answering system. Question Answering (QA) systems are being increasingly finding use in all walks of life. Domain specific QA systems address the task of automatically answering a question posed in natural language and generally deal only

[1] We refer to the latitude and longitude as the spatial information and temporal refers to the time of the query in terms of time of the year etc.

with questions in a specific domain (for example, a QA system for medicine is not expected to answer a query related to automotive maintenance). In many cases domain specific QA systems exploit domain-specific knowledge frequently formalized in ontologies (WikipediaQA, 2012) to function. An ontology formally represents the domain knowledge as a set of concepts, and the relationships between them. These concepts and the relationship between them can be exploited to not only describe the domain but also enable extraction of relationship between different entities within that domain (WikipediaOnt, 2012; Xie et al., 2008). Generally ontologies, especially in the area of agriculture, may be unable to provide personalized level of information. More recently (Bansal and Malik, 2011) provide a framework for crop production life cycle which not only provides relevant but contextual relevant as well as scientifically correct information. However, in an ideal situation a QA system (assisted by a suitable ontology) assumes that the query posed is perfect in the sense of completeness of the information in the query. In reality, the queries are often incomplete and fuzzy. In the absence of an ontology being able to assist address such fuzzy queries would end up making the QA system unable to address the query and/or prone to erroneous answers.

In this paper, we first formulate a multimedia query answering problem and then propose a reconfigurable ontology that assists in answering the multimedia query. The rest of the paper is organized as follow, in Section 2 we formulate the problem and provide a solution approach in Section 3. We give some experimental results in Section 4 and conclude in Section 5.

2 PROBLEM FORMULATION

Given a multimedia query consisting of text, image, video, speech along with spatial and temporal information of the past and future (predicted). We try to address the problem of how to answer the query precisely?

For the sake of clarity, we assume the query to be related to agricultural domain with the understanding that this is in no way restricted to the agricultural domain specifically. Let q_t be the query posed at time t and let q_t in addition to the query text consist of

a Spoken data, S_t, at time t (spoken query in natural language),

b image data, I_t at time t (image or the photograph of the crop related to the query),

c video data, V_t at time t (video of the crop or a farming procedure related to the query),

d spatial location L (the location from where the query originated),

e farmer data P (details of the farmer in terms of land holding),

f farm data F_k for $k \in [t_0, t]$ for location L (details of the farm like agricultural operations done, pesticides and fertilizers used to date),

g weather data W_m for $m \in [t_0, t] \cup]t, t+\delta]$ for location L (like temperature, rainfall obtained from sensors, both past and predicted),

Here δ is the time interval for which prediction data is available and t_0 is some initial finite start time since when information is available. So the spatio-temporal multimedia query q_t can be represented as

$$q_t = (S_t, I_t, V_t, P, L, \{F_k\}_{k=t_0}^{t}, \{W_m\}_{m=t_0}^{t+\delta}), \quad (1)$$

now the problem is that of finding a QA system $\mathcal{F}$ which can use the knowledge $\mathcal{K}$ to identify an answer $\mathcal{A}_t$, namely,

$$\mathcal{A}_t = \mathcal{F}(q_t / \mathcal{K}) \quad (2)$$

For the sake of making this problem tractable, we assume that it is indeed possible to represent the query in multimedia in the form of text. For example,

$$S_t \xrightarrow{\text{speech recognition}} \mathcal{T}_{S_t}$$

where the spoken query S_t can be converted into text (say into $\mathcal{T}_{S_t}$) using an automatic speech recognition engine (Imran and Kopparapu, 2011), and

$$I_t \xrightarrow{\text{image interpretation}} \mathcal{T}_{I_t}$$

where the image I_t can be converted into text (say $\mathcal{T}_{I_t}$) using image processing followed by pattern matching and image analysis, similarly all the spatio-temporal data can be represented in the form of text. Namely,

$$\mathcal{T}_{q_t} = \Big(\mathcal{T}_{S_t} \cup \mathcal{T}_{I_t} \cup \mathcal{T}_{V_t} \cup \mathcal{T}_{\{F_k\}_{k=t_0}^{t}} \cup \mathcal{T}_P \cup \mathcal{T}_L \cup \mathcal{T}_{\{W_m\}_{m=t_0}^{t+\delta}}\Big) \quad (3)$$

Now the problem (2) is one of given a set of text terms, $\mathcal{T}_{q_t}$, derived from different forms of multimedia data, find an answer $\mathcal{T}_{\mathcal{A}_t} (= \mathcal{A}_t)$ given the domain knowledge $\mathcal{T}_{\mathcal{K}} (= \mathcal{K})$, namely,

$$\mathcal{A}_t = \mathcal{F}(\mathcal{T}_{q_t} / \mathcal{K}) \quad (4)$$

2.1 Multimedia Query ($\mathcal{T}_q$)

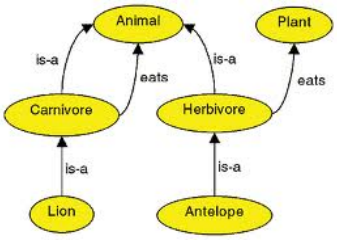
A typical query is shown in Table 1. The first three data items (S_t, I_t, V_t) are sent by the person posing the query and the rest are obtained automatically, for example W_t is obtained from the sensors on the field (TCS, 2012). Note that $\mathcal{T}_{q_t}$ is the union of all the derived text in Table 1. Namely, $\mathcal{T}_{q_t} =$

Data	Sample	Derived Text ($\mathcal{T}_*$)
S_t		I have been having white spots on my grapes what could that be?
I_t		Black Grapes have mild powdery spots visible
V_t	None	–
L	Automatic	12 km north of Nasik
P	Automatic	Ramlal is a knowledgeable farmer, but poor
t	Query Time	Harvesting season in Nasik
F_t	Automatic	Used new seed variety; sprayed Potassium last week
W_t	Automatic	Warmer than usual for this time of the year
W_{t-}	Past Data	Has been cold in the last two days
W_{t+}	Predicted Data	Likely to rain with probability 0.9 in 24 hours

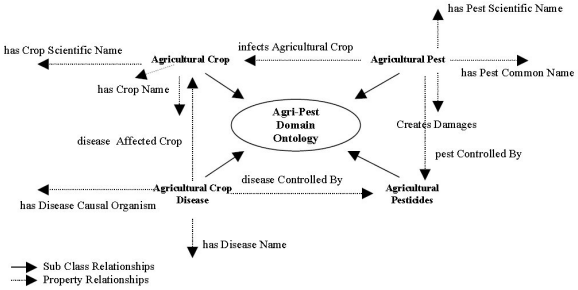
Table 1: Typical Multimedia Query

$$\left\{\begin{array}{l}\text{I have been having white spots on my ... that be?}\\ \text{Black Grapes have mild powdery spots visible}\\ \text{12 km north of Nasik}\\ \text{Ramlal is a knowledgeable farmer, but poor}\\ \text{Harvesting season in Nasik}\\ \text{Used new seed variety; sprayed Pottas .... week}\\ \text{Warmer than usual for this time of the year,}\\ \text{Has been extremely cold in the last two days,}\\ \text{Likely to rain with probability 0.9 in 24 hours}\end{array}\right\}$$

Note that in reality the process of converting a multimedia input (especially, S_t, I_t, V_t) is not very accurate. For example conversion of $S \rightarrow \mathcal{T}_S$ requires a speech recognition of a naturally spoken query by the user. The automatic speech recognition (ASR) accuracy is very poor for resource deficient languages especially when the system has to work in a speaker independent mode. The same holds good for converting $I \rightarrow \mathcal{T}_I$ or $V \rightarrow \mathcal{T}_V$ which would require significant amount of image and video analysis respectively. These areas, namely, speech recognition, image interpretation (Deruyver et al., 2009) and video interpretation (Selman et al., 2011) are being actively addressed by researcher around the globe. For the purpose of our discussion we will assume that the process of converting multimedia data into the corresponding text is error free.



(a) Basic ontology representation.



(b) A agri-pest domain ontology (Urs and Angrosh, 2007) .

Figure 1: Domain ontology representation.

2.2 Knowledge ($\mathcal{T}_\mathcal{K}$)

The domain knowledge, $\mathcal{T}_\mathcal{K}$, is essentially an ontology which by definition is the formal representation of the knowledge in terms of a set of concepts within a domain, and the inter relationships between them (Wikipedia, 2012). A sample $\mathcal{K}$ is given in Fig. 1. While 1(a) shows a sample generic ontology, Fig. 1(b) shows a specific agri-pest domain ontology. In a very loose sense, $\mathcal{K}$ can be looked upon as a graph, where the nodes are the concepts related to the domain (example, `harvest`, $\cdots$) which are related to the entities (for example, `paddy`, $\cdots$) and the edge connect the concepts and the entities. However, the edges can be associated with a strength, in a weak sense, a probability measure that captures the degree

of association between the entity and the concept. As we will show in the next section, these edge weights in an ontology are dependent on the spatio-temporal information in a multimedia query.

2.3 Query Answering ($\mathcal{T}_{\mathcal{A}}$)

The problem is now that of solving (4), namely finding, a $\mathcal{F}$, such that a $\mathcal{T}_{\mathcal{A}}$ can be identified for the query $\mathcal{T}_q$ using $\mathcal{K}$ or alternatively, finding a set of $\mathcal{T}_{\mathcal{A}}$ that are ranked in a way to best address $\mathcal{T}_q$. Note that $\mathcal{F}$ could be, typically a Question Answering System. Typically, one would expect $\mathcal{T}_{\mathcal{A}}$ as

> *Spray insecticide at* 1 *ml per* 1 *litre of water after* 48 *hours.*

in response to the multimedia query q_t.

3 SOLUTION APPROACH

This problem formulation is in the form of a typical Question Answering (QA) system (Kopparapu et al., 2007). A natural language text query is the input to the QA system and the system is to identify a relevant set of possible answers corresponding to the query from a set of answers that is already known to the system.

However, in this paper, we adopt a slightly different approach to address this problem. We visualize the answer to lie in the structure of $\mathcal{K}$ itself; however unlike the typical ontology, we use a reconfigurable $\mathcal{K}$ (Pande and Kopparapu, 2011) which configures itself based on the query there by leading to the correct answer selection.

3.1 Reconfigurable $\mathcal{K}$

The central idea is to construct $\mathcal{K}$ dynamically, the dynamism being triggered by the nature of the query, the time of the query, the origin of the query, etc. The idea being that a dynamic $\mathcal{K}$ captures the context more precisely and hence is able to assist the QA system point to an appropriate answer quickly and reliably. In essence $\mathcal{K}$ captures the environmental conditions that exists at the farmer's location and this enables answering the spatio-temporal multimedia query. Essentially $\mathcal{K}$ dynamically adjusts itself so that it is most relevant to the posed query. This dynamism assists the QA system to perform with higher precision and hence imitating an expert to identify the most *relevant* answer. As mentioned earlier, $\mathcal{K}$ can be looked upon as a graph, where the nodes are the concepts related to the domain (example, `harvest`, ...)

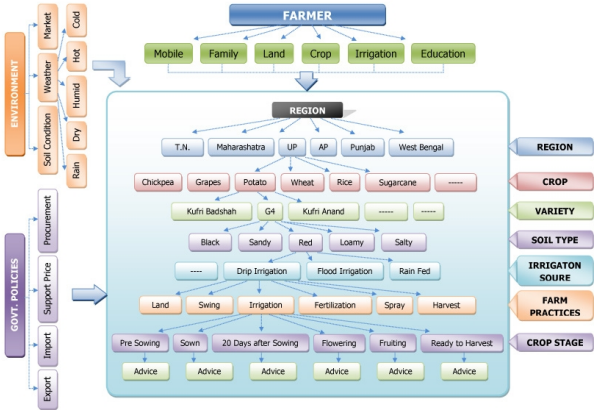


Figure 2: A sample $\mathcal{K}$

which are related to the entities (example, `paddy`, ...) and the edge connecting the concepts and the entities are associated with a strength, in a weak sense, a probability measure. This strength of the edge connecting the concept and the entity is dependent, for example, on the time of the year (season) when the query was asked. For example all the entities connecting to the concept `harvest` would be fully connected (probability 1) in the actual harvesting season (say if the query came in September) while all connections to the concept `harvest` would be broken (probability 0) if the same query were to come in say December (non-harvesting season). Similarly, the connectivity between the concept `harvest` and the entity `paddy` could be 0 for a query coming from a non paddy growing region (say Punjab) and 1 when the query comes from a paddy growing region (say West Bengal). In this particular example, the strength of the connectivity between the concept `harvest` and the entity `paddy` would be the product of all these conditional probabilities, in this example, season and region. Namely the strength of the edge between `harvest` and `paddy` would be $P(\text{season}) \times P(\text{region})$. As another example, the association between `spray` and `insecticide` would be based on the probability that it would rain in the next couple of days. A higher probability of rain would decrease the association while a prediction of no rain would strengthen the association. A typical $\mathcal{K}$ for agricultural practices is shown in Fig. 2.

4 EXPERIMENTAL RESULTS

We handcrafted $\mathcal{K}$ as shown in Fig. 2 with the help of agricultural experts. We extracted about 100 real queries that the farmers has asked and the actual answers given by the human expert on our platform (TCS, 2012). The queries consisted of multimedia in-

formation which we converted into text manually by listening to the audio query and by looking at the images associated with the query. In all the cases the response of the expert was a text message. For each of the query, we also extracted the environmental condition information and also the time of the query etc from logs. In almost 90 % of the cases we were able to derive the response from $\mathcal{K}$. For example, an actual response of the expert say "*Spray insecticide at* 1 *ml per* 1 *litre of water after* 48 *hours.*" gave us "*Do nothing*" because the concept rain removing the connection between insecticide and activity of spraying. The experimental results are encouraging and we are in the process of studying further the reconfigurability of $\mathcal{K}$ so as to enable multimedia question answering.

5 CONCLUSIONS

Multimedia queries are common in a variety of domain more so in agriculture domain because of the need for the expert to ascertain several environmental conditions to be able to give a reliable answer. Very often the answers are different for the same query just because the query came from a certain region or at a certain time of the year. In this paper, we first formulated the problem as a multimedia query answering system and then presented a self configuring $\mathcal{K}$ based on the query itself, this is the main contribution of the paper. Preliminary results, shows that this approach is usable in addressing multimedia queries, provided that the multimedia data is converted into text data without any error. We plan to study the problems that can crop up when errors are introduced when multimedia query is converted into text.

REFERENCES

Bansal, N. and Malik, S. (2011). A framework for agriculture ontology development in semantic web. In *International Conference on Communication Systems and Network Technologies*, pages 283 – 286.

Deruyver, A., Hodé, Y., and Brun, L. (2009). Image interpretation with a conceptual graph: Labeling oversegmented images and detection of unexpected objects. *Artif. Intell.*, 173(14):1245–1265.

Imran, A. and Kopparapu, S. (2011). Building a natural language hindi speech interface to access market information. In *Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2011 Third National Conference on*, pages 58 –61.

Kopparapu, S. K., Srivastava, A., and Rao, P. V. S. (2007). Minimal parsing key concept based question answering system. In Jacko, J. A., editor, *HCI (3)*, volume 4552 of *Lecture Notes in Computer Science*, pages 104–113. Springer.

Lopez, V., Uren, V. S., Motta, E., and Pasin, M. (2007). Aqualog: An ontology-driven question answering system for organizational semantic intranets. *J. Web Sem.*, 5(2):72–105.

Pande, A. and Kopparapu, S. K. (2011). A self configuring knowledge base representation, indian patent number 3294/mum/2011.

Selman, J., Amer, M. R., Fern, A., and Todorovic, S. (2011). Pel-cnf: Probabilistic event logic conjunctive normal form for video interpretation. In *ICCV Workshops*, pages 680–687. IEEE.

TCS (2012). http://www.tcs.com/offerings/technology-products/mkrishi/pages/default.aspx.

Urs, S. R. and Angrosh, M. A. (2007). Ontology-based knowledge organization systems in digital libraries: A comparison of experiments in owl and kaon ontologies. `www.vidyanidhi.org.in/shaliniurs_files/icdl.pdf`.

Wikipedia (2012). http://en.wikipedia.org/wiki/ontology.

WikipediaOnt (2012). http://en.wikipedia.org/wiki/ ontology_computer_science.

WikipediaQA (2012). http://en.wikipedia.org/wiki/ question_answering.

Wu, G.-L., Su, Y.-C., Chiu, T.-H., Hsieh, L.-C., and Hsu, W. H. (2011). Scalable mobile video question-answering system with locally aggregated descriptors and random projection. In *Proceedings of the 19th ACM international conference on Multimedia*, MM '11, pages 647–650, New York, NY, USA. ACM.

Xie, N., Wang, and Yang, Y. (2008). Ontology-based agricultural knowledge acquisition and application. I 258:349– 357.

SESSION

DATABASES, BIG DATA, VISUALIZATION, COMPRESSION, DATA PRIVACY, AND TOOLS

Chair(s)

TBA

DATABASE QUERYING USING FUZZY SEMANTICS

Punita Thammareddy and Devinder Kaur
Electrical Engineering and Computer Science
University of Toledo
Toledo, Ohio

Abstract— *Search engines play a vital role in daily and corporate life starting from World Wide Web search engine to a simple library search engine. A good search engine should retrieve relevant data even with imprecise input or data as the users would generally use natural language while searching for something and if the search is not modeled correctly, it would retrieve irrelevant data. To resolve this issue fuzzy logic is inferred into the conventional database. This paper illustrates with an example on how fuzzy logic is implemented on MySQL database. With the development of a user interface, this example can as well be served as a search engine related to weather, climate and cities.*

Keywords: Databases, Query, Fuzzy Semantics.

1 INTRODUCTION

The modern industrialized society we live in makes us more and more dependent on ever increasing amounts of information to lead a peaceful and comfortable routine life or to cope up or compete with the competitive corporative life. The solution for this is a search engine which in simple terms is any software program that searches for site, based on the words users have keyed in, that have to do with the subject of their interest. But except for few search engines like Google are not successful because of the hundreds of results these search engines generate are not all relevant to the search criteria. Generally, the users use natural language for searching which can result in an imprecise and/or vague and/or uncertain and/or inconsistent and/or ambiguity criteria like "rather cold city near Toledo". So the major issue these search engines need to address in almost every discipline of data processing is the evaluation of data in terms of relevance criteria such as correctness, completeness, conciseness, confidence, quantity, etc.

How can we design a data mining model like a search engine which gives result set with satisfactory degree of relevance? Clearly, conventional database techniques don't provide convincing results. To resolve this issue a concept of "Fuzzy Logic" is applied to database technique. In this paper, the basic concepts of fuzzy logic, how it can be implemented on database and few SQL queries with fuzzy concept are modified slightly to achieve for a search engine for the cities based on their weather conditions like "cold city", "snowy city", "wet cities" etc. The paper is structured as follows. In Section II, relevant literature review is presented. In Section III, implementation of fuzzy logic on a database with few example queries and their result sets are discussed. Section IV provides the conclusion and guidelines for future work.

2 LITERATURE REVIEW

Fuzzy logic was first introduced by L. A. Zadeh [1]. Fuzzy logic is a multi-valued logic that deals with reasoning rather than the traditional logic which deals with exact and fixed values. The truth value of a fuzzy variable ranges between (0 and 1), (true and false), or (yes and no), or (high and low), etc. Furthermore, the membership degree, in other words "degree of truth" can be determined from the membership function of a fuzzy variable. L. A. Zadeh [1] has proposed several different types of membership functions and can be used depending on the type of the fuzzy attribute. Authors of [2-6] have discussed about different models of fuzzy databases and different ways to implement imprecise queries using natural language.

Zheng Pei, Yang Xu, Da Ruan, Keyun Qin, Dan Meng [7-8] discussed on how to optimize the complex linguistic queries using genetic algorithms. Daniel Pilarski [9] describes about the Quantirius, an interactive system supports data mining and linguistic summaries in a database, whose assessment of validity degrees are realized by Zadeh's fuzzy logic based calculus. Janusz Kacprzyk, Sławomir Zadro[10] advocate Zadeh's prototype-centered approach and illustrated it by implementing the technique on sales database for a computer retailer.

3 EXPERIEMNT AND RESULTS

The average weather conditions varies widely from place to place and if anyone is planning or willing to move temporarily or permanently, the weather conditions of that places play a vital role in the decision making. In this paper, the average monthly temperature, average annual precipitation, and average annual snowfall for 100 cities in the United States is downloaded from the website: http://www.infoplease.com/ipa/A0762183.html and Cloudiness, Average Wind Speed, and Average Relative Humidity of selected cities is downloaded from the website:

http://www.census.gov/compendia/statab/cats/geography_environment/weather_events_and_climate.html.

This experiment is based on the approach proposed by Bosc [11] show in Figure 1. As shown in Figure 1, the fuzzy query process is done by a transformation procedure located on top of the existing DBMS. The translation mechanism generates a procedural evaluation program and determines the expressions which are used to compute the membership degrees. In this experiment, the transformation procedures applied to MySQL Server are illustrated.

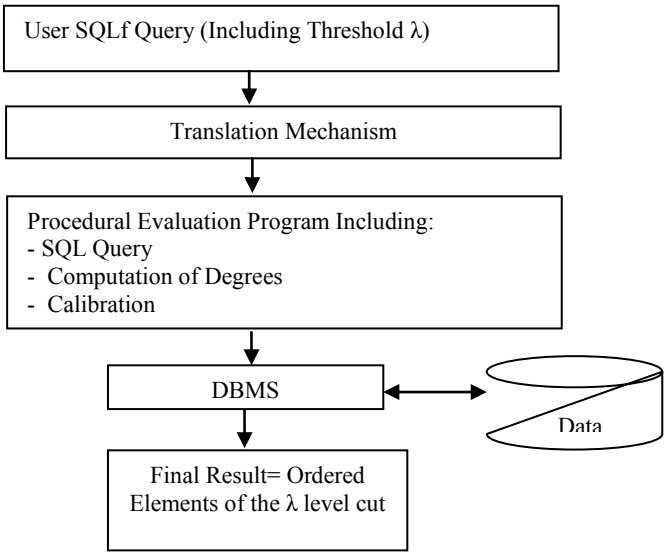


Figure 1: Flow chart based on Bosc's approach.

Source: Bosc P. and Pivert O., "SQLf Query Functionality on Top of a Regular Relational Database Management"

This implementation of fuzzy semantics on MySQL database is done in two ways.

1.) Calculating the membership degree with the predefined boundary values, putting these values into a table and then retrieving the data respectively.

2.) Calculating the membership degree based on the inputs given by the user while retrieving the data simultaneously.

Part 1: As the downloaded data consists average monthly temperatures, average annual precipitation, and average annual snowfall each of which can have a set of membership functions like temperature can have membership function namely "Cold" and "Hot", average annual precipitation can have "LessPrec" and "MorePrec" and average annual snowfall can have "LessSnow" and "MoreSnow". The membership functions of these attributes are shown in figure 2.

Both the "Cold" and "Hot" membership functions are trapezoidal and they are calculated using the following formula as shown in equation 1 that is given in [12]:

$$T(x) = \begin{cases} 0 & if\ (x<a)and\ (x>d) \\ \frac{(x-a)}{(b-a)} & if\ (x>a)and\ (x<b) \\ 1 & if\ (x>b)and\ (x<c) \\ \frac{(d-x)}{(d-c)} & if\ (x>c)\ and\ (x<d) \end{cases} \qquad (1)$$

Source: Galindo, J., Urrutia, A., & Mario, P. (2006). *Fuzzy databases: modeling, design, and implementation.*

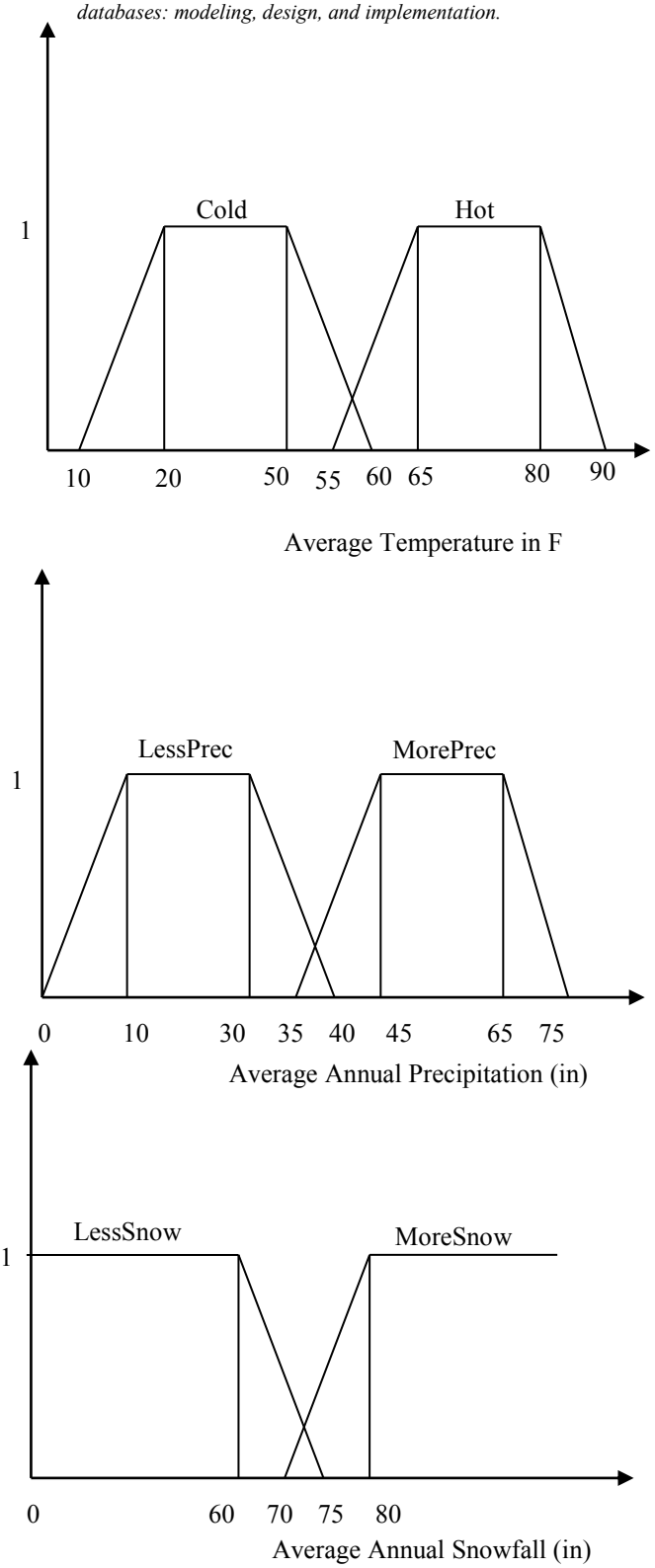


Figure 2: Membership Functions of the attributes.

These membership functions are created accordingly using SQL syntax and their boundary values as shown in the figure 2, one of which is given below.

```
CREATE DEFINER=`root`@`localhost`
FUNCTION `Cold`(x  float(10)) RETURNS float
Begin
    declare a integer;
    declare b integer;
    declare c integer;
    declare d integer;
    declare res float(10);
     set  a=10;
    set b=20;
    set c=50;
    set d=60;

  if ( x <= a or x >= d) then
 set res=0;
  elseif (x>a and x<b) then
 set res= (x-a)/(b-a);
  elseif (x>=b and x<=c) then
 set res=1;
  else
 set res= (d-x)/(d-c);
  END  if;
return res;
 END
```

The above is the syntax for membership function "cold". The rest of the membership functions are *"Hot", "LessPrec", "MorePrec", "LessSnow", and "MoreSnow"*. Now, using these membership functions and the data available membership degree of each tuple is calculated and Inserted into a new table. The following SQL query calculates the membership degree and inserts it into a new table *TempSnow_membership*.

```
INSERT INTO TempSnow_membership()
        (SELECT city,state,
    Cold(AvgMonthlyTemp_F),
    Hot(AvgMonthlyTemp_F),
    MoreSnow(AvgAnnualSnowfall_inch),
    LessSnow(AvgAnnualSnowfall_inch),
    MorePrec(AvgAnnualPrecipitation_inch),
    LessPrec(AvgAnnualPrecipitation_inch) FROM
    (SELECT * FROM TempPrecipitationSnowfall) r
    );
```

After populating, this table looks like Table 1.

Table 1: This is how the populated table data looks like.

City	State	Cold	Hot	MoreSnow	LessSnow	MorePrec	LessPrec
Albany	New York	1	0	0	0.706667	0.36	0.14

Now, to get the list of cities and states which are "Hot" with a threshold of 0.8 can be retrieved by the following SQL Query:

```
SELECT city,state,hot,cold FROM
TempSnow_membership where Hot>0.8;
```

The results for the above query are in Table 2. The same way by changing the threshold we can retrieve the list based on the membership function.

Part 2: For this part, the task of retrieving the list of cities whose average temperatures are "around" the given temperature with a given threshold value is considered.

For this step, a guassian function is used as the membership function is shown in Figure 3, has the memebership function as follows:

$$G(x) = exp\ (-k*(x-m)^2) \qquad (2)$$

where k>0 and greater the value of k, narrower the bell. m is center and k is the standard deviation.

Here, k is 1 and m is to be defined by the user.

Source: Galindo, J., Urrutia, A., & Mario, P. (2006). *Fuzzy databases: modeling, design, and implementation.* Hershey: Idea Group Inc.

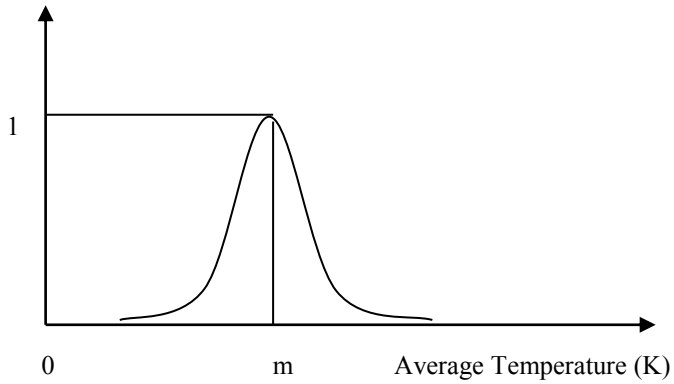


Figure 3: Gaussian Fuzzy Set

The membership functions *"Cloudy", "Windy", "Humidity"* and *"Temp"* are similar and can be defined as follows:

```
CREATE DEFINER=`root`@`localhost` FUNCTION
    `Temp`(x  float(10), m float(10)) RETURNS float
BEGIN
declare k integer;
declare res float;
 set  k=1;
 set res= EXP(-k*(x-m)*(x-m));
  return res;
 END;
```

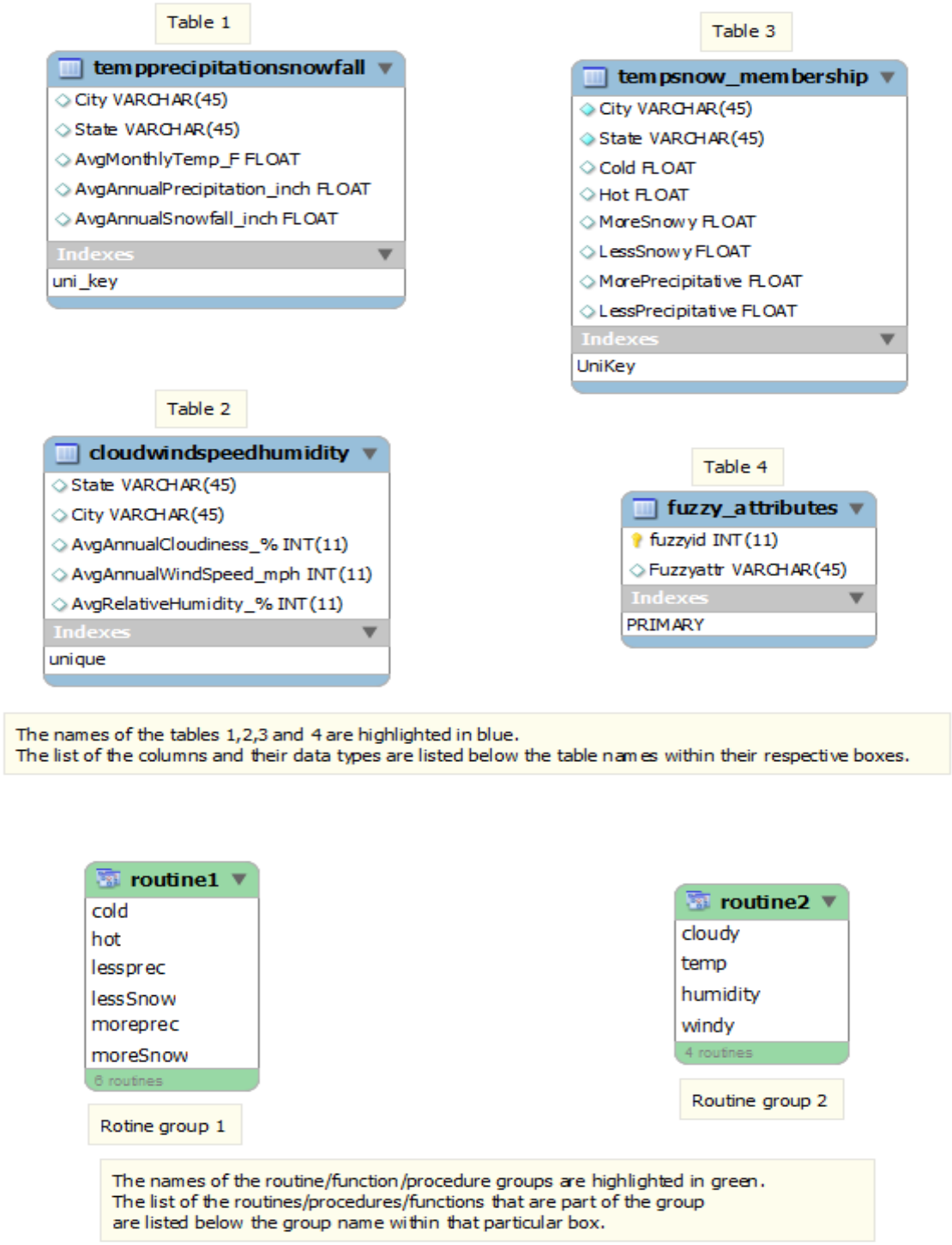


Figure 4: Shows the Enhanced Entity Relationship (EER) Model of the database.

The Enhanced Entity Relationship (EER) Model shown in Figure 4 consists of four tables and ten routines. The tables *TempPercipitationSnowfall* and *CloudWindspeedHumidity* consist of the data downloaded from the above mentioned websites. The *routine1* is the group of routines that are used to calculate membership functions as described in part 1. The table *TempSnow_membership* consists of values of membership functions that are calculated from the routines from *routine1* group. The *routine2* is the group of routines that are used to calculate membership functions as described in part 2. The table *Fuzzy_attributes* contains the list of all the fuzzy attributes used in this database.

Table 2: The result set of hot cities with threshold=0.8

city	state	Cold	Hot
Austin	Texas	0	1
Baton Rouge	Louisiana	0	1
Charleston	South Carolina	0	1
Columbia	South Carolina	0	0.8375
Dallas-Ft. Worth	Texas	0	1
El Paso	Texas	0	0.9475
Honolulu	Hawaii	0	1
Houston	Texas	0	1
Jackson	Mississippi	0	0.855
Jacksonville	Florida	0	1
Las Vegas	Nevada	0	1
Long Beach	California	0	1
Los Angeles	California	0	0.8525
Miami	Florida	0	1
Mobile	Alabama	0	1
Montgomery	Alabama	0	0.9525
New Orleans	Louisiana	0	1
Phoenix	Arizona	0	1
San Antonio	Texas	0	1
San Diego	California	0	0.9725
Savannah	Georgia	0	1
Tampa	Florida	0	1
Tucson	Arizona	0	1
Vero Beach	Florida	0	1

The SQL query to retrieve the list of cities and states whose average temperature is *"around 58 K"* with a threshold of 0.1 is as follows:

SELECT city, state, AvgMonthlyTemp_F, *Temp*(AvgMonthlyTemp_F,58) as membership from TempPrecipitationSnowfall where *Temp*(AvgMonthlyTemp_F,58)>=0.1

The above query results are in Table 3 below.

Table 3: The result set with temperatures "around 58K with threshold=0.1"

city	state	AvgMonthly Temp_F	membership
Albuquerque	New Mexico	56.775	0.22299158573150635
Knoxville	Tennessee	57.975	0.9993751049041748
Louisville	Kentucky	56.575	0.13125374913215637
Nashville	Tennessee	58.575	0.7184739112854004
Norfolk	Virginia	59.425	0.13125374913215637
Raleigh	North Carolina	59.4	0.14085781574249268
Richmond	Virginia	57.425	0.7184739112854004
San Francisco	California	57.35	0.6554049253463745
Washington	District of Columbia	57.25	0.5697828531265259

The SQL query to retrieve the list of cities and states whose average annual wind speed is *"around 7 m.p.h with a threshold of 0.1*" is as follows:

SELECT city, state, AvgAnnualWindSpeed_mph, *Windy*(AvgAnnualWindSpeed_mph,7) as membership from CloudWindSpeedHumidity where *Windy*(AvgAnnualWindSpeed_mph,7)>=0.1;

The above query results are shown in Table 4 below.

Table 4: The result set with average wind speeds "around 7 mph with threshold=0.1"

city	state	AvgAnnual WindSpeed_mph	Windy
Juneau	Alaska	8	0.3678794503211975
Phoenix	Arizona	6	0.3678794503211975
Little Rock	Arkansas	8	0.3678794503211975
Los Angeles	California	8	0.3678794503211975
Sacramento	California	8	0.3678794503211975
San Diego	California	7	1
Hartford	Connecticut	8	0.3678794503211975
Jacksonville	Florida	8	0.3678794503211975
Louisville	Kentucky	8	0.3678794503211975
New Orleans	Louisiana	8	0.3678794503211975
Jackson	Mississippi	7	1
Reno	Nevada	7	1
Concord	New Hampshire	7	1
Charlotte	North Carolina	7	1
Raleigh	North Carolina	8	0.3678794503211975
Columbus	Ohio	8	0.3678794503211975
Portland	Oregon	8	0.3678794503211975
Columbia	South Carolina	7	1
Nashville	Tennessee	8	0.3678794503211975
Houston	Texas	8	0.3678794503211975
Richmond	Virginia	8	0.3678794503211975
Charleston	West Virginia	6	0.3678794503211975
San Juan	Puerto Rico	8	0.3678794503211975

The SQL query to retrieve the list of cities with temperatures *"around 58K"* and with wind speed *"around 7 mph"* both with threshold=0.1 is as follows:

```
SELECT r.city, r.state, AvgMonthlyTemp_F,
temp, AvgAnnualWindSpeed_mph, windy from
    (SELECT city, state, AvgMonthlyTemp_F,
Temp(AvgMonthlyTemp_F,58) as temp from
TempPrecipitationSnowfall where
Temp(AvgMonthlyTemp_F,58)>=0.1)r JOIN
    (SELECT city, state,
AvgAnnualWindSpeed_mph,
Windy(AvgAnnualWindSpeed_mph,7) as windy from
CloudWindSpeedHumidity where
Windy(AvgAnnualWindSpeed_mph,7)>=0.1)s on
r.city=s.city and r.state=s.state;
```

The above results are shown in Table 5 below.

Table 5: The result set with "temperatures around 58K" and "wind speeds around 7 mph"

city	state	AvgMonthly Temp_F	Temp	AvgAnnual WindSpeed _mph	Windy
Raleigh	North Carolina	59.4	0.140858	8	0.367879
Nashville	Tennessee	58.575	0.718474	8	0.367879

4 CONCLUSION AND FUTURE STUDY

The basic concept of fuzzy logic has been implemented on non-fuzzy DBMS i.e. MySQL database. Now, we can only retrieve the data related to weather conditions alone. This can be further extended to GIS by adding the location co-ordinates of the cities so that their distance also can be calculated as a fuzzy set. For example, "cold or rainy cities near or around Toledo" would be a reasonable fuzzy search engine for weather related information. Also, a User Interface should be created which would take the "natural language query" and convert it to SQL query in order to make it a complete search engine.

5 REFERENCES

[1] L. A. Zadeh, "*Fuzzy Sets*", Information and Control, vol. 8, no.3, pp 338-353, June 1965.

[2] Qinzheng Kong, Graham Chen, " On deductive databases with incomplete information".

[3] Z. M. MA+ AND LI YAN, "A *Literature Overview of Fuzzy Database Models*", College of Information Science and Engineering, Northeastern University, Shenyang, 110004

[4] Miroslav Hudec, "An *Approach to Fuzzy Database Querying, Analysis and Realisation*", INFOSTAT - Institute of Informatics and Statistics, Bratislava, Slovakia

[5] Ignacio J. Blanco, Olga Pons, Jose M. Serrano, and M. Amparo Vila, "*Deduction in a GEFRED database using Datalog*"

[6] L. Ughetto, W.A. Voglozin, N. Mouaddib, "*Database querying with personalized vocabulary using data summaries*", Fuzzy Sets and Systems, Volume 159, issue 15 (August 1, 2008), p. 2030-2046

[7] Zheng Pei, Yang Xu, Da Ruan, Keyun Qin, "*Extracting complex linguistic data summaries from personnel database via simple linguistic aggregations*", Information Sciences, Volume 179, issue 14 (June 27, 2009), p. 2325-2332

[8] Dan Meng, Zheng Pei, "*Extracting linguistic rules from data sets using fuzzy logic and genetic algorithms*", Neurocomputing, Volume 78, issue 1 (February 15, 2012), p. 48-54

[9] Daniel Pilarski, "*Linguistic Summarization of Databases with Quantirius: a Reduction Algorithm for Generated Summaries*", International Journal of Uncertainty Fuzziness and Knowledge Based Systems , Volume: 18, Issue: 3, Pages: 305-331.

[10] Janusz Kacprzyk , Sławomir Zadro, " Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools".

[11] Bosc P. and Pivert O., "SQLf Query Functionality on Top of a Regular Relational Database Management," in Proceedings of Knowledge Management in Fuzzy Databases, Heidelberg, pp. 171'190, 2000.

[12] Galindo, J., Urrutia, A., & Mario, P. (2006). *Fuzzy databases: modeling, design, and implementation.* Hershey: Idea Group Inc.

The Centinel Data Format: Reliably Communicating through Time and Place

Clarence Lehman[1], **Shelby Williams**[2], **and Adrienne Keen**[3]
[1]University of Minnesota, 123 Snyder Hall, 1474 Gortner Avenue, Saint Paul, MN 55108, USA
[2]University of Minnesota, 100 Ecology, 1987 Upper Buford Circle, Saint Paul, MN 55108, USA
[3]London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK

"A library book lasts as long as a house, for hundreds of years."
—Thomas Jefferson, 1821

Abstract— *A common experience among scientists and engineers is storing and sharing data, the capacity for which has advanced immensely since laboratory notebooks were only paper and ink. However, since that time, the sustainability of data has decreased. Even though our digital data should be safer and more secure than ever, a continuing cascade of obsolescence in computer media and software can actually make it less so. Here we outline an ensemble of free tools and techniques that we call "Centinel," designed to manage, communicate, and archive digital datasets. Rather than embedding error-correcting codes as part of the computer media, Centinel exposes them and places them with the data and metadata. Thus even printed copies of the data form reliable storage media that can last indefinitely without intervening attention. Centinel complements standard methods for data sustainability, such as data migration. Unified approaches, as we outline here, benefit reliability and longevity of data.*

Keywords: database, data archive, data longevity, data reliability, error correcting codes

1. Introduction

In 1815 began one of the largest scientific data collection projects ever launched [1]. Legions of surveyors walked regularly spaced transects along 2,500,000,000 meters of the Louisiana Territory, recording the biological species, geographic locations, and diameters of selected trees near periodic sample points—plus other information on soils, vegetation, and boundaries of wetlands. For almost a century the survey continued. Now, another century after the last data were recorded, the results form one of the most visible efforts ever, organizing the rural landscape into square sections along those transects. The results also form one of the best preserved and widely available datasets ever. Think of which present datasets, in your personal experience, are guaranteed to be extant and usable well into the 22nd century.

A large part of the reason the survey data survived was that it was recorded on paper and protected at many different governmental sites. In the meantime, technology changed immensely. Computers emerged and increased in capacity so relentlessly that the Library of Alexandria's ancient charge of organizing and cataloging all human knowledge began to draw within reach. Global access to digital data can make that knowledge available to all. Large-scale private enterprises are aiming at this goal, but individuals in academia and industry are established sources of knowledge and therefore have a special role in achieving this.

Here we are addressing that role—of scientists, engineers, and others who collect empirical data, share it, and want to preserve it for the future. In this report we explain how digital computer techniques of today combine naturally with paper methods of prior centuries to create a form of digital storage that can reliably persist into future centuries and improve electronic processing today.

2. What Centinel is and is not

The general topic that Centinel addresses has been long discussed (e.g., [2] [3] [4] [5] [6] [7] [8] [9]) and a complete solution is not yet available. Centinel combines the words "century" and "sentinel," guarding data for extended periods. One goal for Centinel is to ensure that the digital data it encodes will be accessible in a century or more, without the need for care and intermediate steps by humans. A second goal is to protect data over a shorter term, from the time of initial creation to the time of final processing. Centinel works by (1) keeping all metadata with the data, (2) protecting data with line-by-line error correcting codes, (3) providing a format easily readable by humans as well as computers and scanners, (4) supporting a reliable digital format that works on any media, including paper and verbal communications, to protect data from unintentional alteration, and (5) supplying an extensible, self-defining format with accompanying tools that help computer programmers know that the data entering their programs are correct. Centinel is an approach to data management, but also a set of basic computer utilities for writing, reading, editing, separating, joining, ordering, and aligning data. It avoids structures that are error prone

```
6674844762232577  Keyword SpAbbr: Abbreviations for species names. Abbreviations contain the
0629561874138616    first three letters of the genus name followed by the first three letters
0211050455008008    of the species name. The full species names are recorded with their
5515307245627135    abbreviations in table "species codes" at the end of the chapter.
5915322805104717  Keyword Date: Date species was collected. Format year-month-day.
1453182442695072  Keyword CollID: Unique code assigned to species sample collected.
1382423906566782  Keyword Cover: Estimated canopy cover, in percent. Dashes indicate missing
5953391885352618    data. (See "methods" at the end of the chapter.)
0748783303437946  Keyword HtMax: Maximum height, in meters. Dashes indicate missing data.
0229302812296440    (See "methods" at the end of the chapter.)
0602554115737437  Keyword HtMin: Minimum height, in meters. Dashes indicate missing data.
0229302812296440    (See "methods" at the end of the chapter.)
0000000000000000
1976160343505769  :Site :Code     :SpAbbr  :Date       :CollID   :Cover :HtMax :HtMin
4554847814214755  :1600 :P1600D04 :Abibal  :1989-08-21 :AMB00555 :    - :    5 :    5
2645745581124348  :1600 :P1600D01 :Abibal  :1989-08-21 :AMB00604 :    2 :    1 :    1
1076375677295808  :1600 :R1600EA  :Abibal  :1989-08-24 :AMB00666 :    3 :    1 :    1
2000445884315808  :1600 :R1600EA  :Abibal  :1989-08-24 :AMB00668 :    5 :    6 :    6
0582355170295008  :1600 :R1600EA  :Abibal  :1991-08-05 :AMB01719 :    2 :    2 :    2
1485325476235008  :1600 :R1600EA  :Abibal  :1991-08-05 :AMB01722 :    4 :    6 :    6
4100414960104041  :1600 :R1600EA  :Acerub  :1991-08-05 :AMB01503 :    2 :    2 :    2
5773084583093978  :1600 :P1600B01 :Agrsca  :1989-08-25 :AMB00456 :    3 :    2 :    2
4766066289426272  :1600 :P1600D01 :Amerot  :1991-06-17 :AMB01439 :    2 :    2 :    2
```

Figure 1. *Excerpt of a sample Centinel data file from a large ecological database, with metadata above and error correcting codes called "centinels" at left. Here colons separate columns rather than vertical bars. In the Centinel structure, error detection and correction stays with the data rather than with the computer medium.*

and supports good data management practices, for example as outlined in [10] and [11].

Centinel is not intended to substitute for large-scale interactive databases undergoing continual manipulation, such as in PostgreSQL, MySQL, or Access. It is, however, a good format for long and medium-term retention of such databases, as Centinel format can be readily exported from them through simple utility programs, and conversely, imported through conventional means or by scanning. Nor is Centinel intended as a complete solution to the problem of storing all data at national and international scales (e.g. [12] [13]), but rather as a solution for individual research and development groups to help maintain their data.

The Centinel format shown in Figure 1 supports the movement of data through place and time. A dataset documented sufficiently with complete descriptions as its metadata, and protected with error correcting "centinels," can be transmitted to another researcher in a distant place without separate documentation and time spent explaining the data, or equivalently it can be transmitted forward to another researcher in the distant future. In other words, it can be archived. Instead of error detecting and correcting codes being applied to the storage media, as is the common method today, codes in Centinel are applied to the data themselves, and stay with the data through all media changes. That simple but unusual characteristic fills a gap in existing data methods and provides confidence in the data across distant places and times. Multiple printed copies of the data can be stored throughout the world and scanned with optical character recognition in the remote future. The centinels, checked automatically against the scanned results, are the essential link to data reliability.

As in some other databases, Centinel has multiple equivalent formats, which we call "singular," "columnar," and "mixed." Long lines of data in singular format can extend onto new lines, indented as in Figure 1. Here is a simpler file in singular format:

```
Class:  1
ID:     123
Age:    21
Region: SSA

Class:  1
ID:     47
Age:    7
Region: UK

Class:  2
ID:     723
Age:    70
Region: US
```

Below are the same data in columnar format:

```
| Class | ID  | Age | Region
| 1     | 123 | 21  | SSA
| 1     | 477 |  7  | UK
| 2     | 723 | 70  | US
```

And below is mixed format:

```
Class: 1
| ID  | Age | Region
| 123 | 21  | SSA
| 477 |  7  | UK

Class: 2
| 723 | 70  | US
```

These formats are interchangeable. The choice is a matter of space, readability, and ease of processing. All software written to handle Centinel data should process the three formats equally.

Printed copies of data with error-correcting centinels need not be limited to small data sets. For example, the genome of the fruit fly (*Drosophila melanogaster*), represented with one base-64 symbol for each of its 47 million codons, would require approximately 6000 pages—not absolutely prohibitive to print for an important, expensive dataset. By comparison, the King James Bible is 4.3 million characters, about one-tenth of this genome, and more than one copy of that work has been printed.

3. How Centinel works

Centinel protects data when they are complete and ready to be archived. But it can also be used when the data are first entered, to guard against accidental modifications of datasets undergoing incremental change.

To explain how Centinel works, we must consider what it means for data to be digital. Two properties are essential. First, the data must be represented by "symbols" that have only a finite number of states. Second, the shapes of any two distinct symbols must be separated by a sufficient gap, so that a symbol for one datum does not, except very rarely, degrade into a different symbol for a different datum. Symbols can take various forms—binary 0 and 1 encoded electronically in computer memories are one example of digital data. The Arabic numerals 0–9 printed on paper are another. With these ideas in mind, Figure 2 shows analog versus digital representations of a function, $y = f(x)$.

An analog form on paper could take the form of a graph, Figure 2A. The value on the vertical axis varies smoothly, and can be read to reasonable accuracy with a ruler and a careful eye. However, each time the graph is copied, its accuracy diminishes. The curve becomes successively blurred, the right side may get slightly skewed with respect to the left, and so forth. In contrast, the entire curve in digital form is defined by coefficients, Figure 2B. When this digital version is copied by re-typesetting, it will not degrade, for the individual symbols will be recognized for what they are and reproduced intact. A new font may even change 'x' to 'x', but the meaning of the symbol will remain.

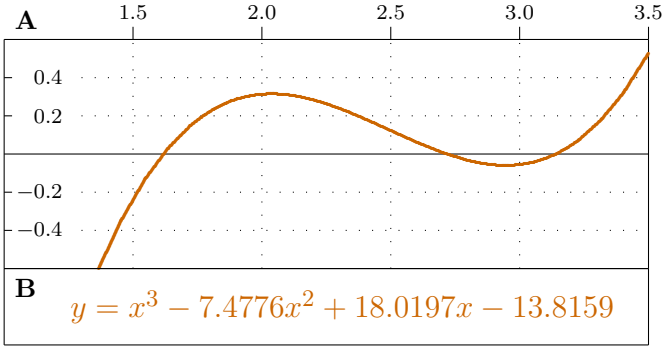


Figure 2. *Non-electronic analog and digital data for the same curve. Printed copies of the digital data (B) will not degrade over time as will the analog version (A) of the same data.*

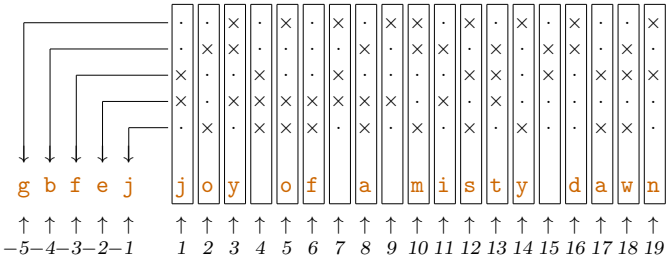


Figure 3. *Error-correcting "centinels" (left) for a 19-character message (right). Each centinel covers a distinct combination of columns, such that any unmatched centinels identify which column is in error and how to correct it. (See code in the appendix for details.)*

Thus digital data are not at all restricted to electronic media, but paper can carry digital data as well, and has done so for millennia. Moreover, some of the most common digital information read by computers today is recorded directly on paper, plastic, metal, and other substrates. The ubiquitous bar code is a case in point, though bar codes are not human-readable as Centinel-protected data are.

A significant separation between symbols in appearance or physical state keeps unavoidable small degradation in information from changing the message, because one symbol does not easily degrade into another. However, separation of symbols is not enough. For highest reliability, error correcting codes must be applied to the digital data to prevent rare alterations of one symbol into another from changing the message, except with negligibly small probability.

Centinel uses a "Hamming code" for arbitrary symbols, a generalization of the original code [14] for binary digits. Such codes we call "centinels," and they appear at the left of each line, at the end of each printed page, and at the end of each file. They can correct any single-symbol error in a line and detect any two-symbol errors. In addition, with high probability they detect multiple-symbol errors, including errors in the centinels themselves.

Each symbol is assigned a small integer and the integers for a given subset of columns are summed. The sum, modulo the number of symbols, is translated back to a symbol, as in columns -1 to -5 of Figure 3. This is repeated for carefully chosen subsets of columns which allow errors to be located and corrected. Then the results are translated to decimal form, as in Figure 1, to mask the actual random combinations of symbols, which by happenstance can spell out any word.

Complete details are in the Centinel algorithms (appendix). These details are part of the metadata and should be included with archived data.

4. Comparison with other approaches

A standard approach to data archiving is a rigorous effort of continually transferring data from old media and old software to new, before the old media and software become completely obsolete—keeping the data "alive" so to speak.

That is called "migration" [12]. It is a practical, well-tested method, though it can be labor intensive and susceptible to catastrophic failure.

Successful migration requires a central discipline maintained over long periods. Any lapse in the chain of migration will result in the complete loss of data. Successful migration will be practical for large, well funded data sets. However, for many small data sets, discipline and funding can easily lapse over long periods of time.

Timing is key, as migration must take place while (1) machines that can read the media still exist, (2) programs encoding the information are still operational, and (3) the media and the information stored on it have not deteriorated.

It follows that the best chance of success in data preservation will be for (1) media that require no advanced or specialized machinery to read them, (2) formats that require no complex computer programs to process them, or at worst require the simplest programs that can be described completely in a few pages of text, as in the Centinel algorithm (appendix), and (3) media and encoding methods that will themselves last a century or more. Centinel allows data preservation with a single migration.

A second method is called "encapsulation." Fully successful migration to new media will be worthless if the software that accesses the data ceases to exist. For example, an organization producing software may go out of existence and no other organization may support the old format. This has happened repeatedly in the history of computing. Encapsulation aims to include with the data all software that accesses the data, in a form that can be translated to future machinery. That is, of course, easiest when the corresponding software is as limited as possible.

Two other methods proposed for data archiving are "emulation" and "technology-preservation." In emulation, the complete hardware and software architectures to retrieve the data are migrated forward with the data and "emulated" on the future system. That practice was widespread and successful among mainframe computers in the 1960s, where one generation of computers would emulate the hardware of the generation before. But as computers become increasingly complex in their architecture and operating software, it becomes difficult to make this practical into the indefinite future.

In technology-preservation, the actual hardware and software is preserved, museum-style, along with the data for future access. This is problematic, however, for today's computers are built for the moment, not built to last, and may not even boot up properly after a decade of disuse.

Therefore, emulation and technology-preservation are not related to Centinel, but migration and encapsulation are. Centinel implements encapsulation in the simplest form—under 100 lines of code (appendix)—and with a single migration, creates digital documents that last as long as possible—up to a century or more.

5. Suggestions

In conclusion, we offer the following: (1) To keep electronic data safe, prepare early for archiving. (2) Archive data in the simplest formats possible. (3) Document data to the highest standards. (4) Associate documentation directly with the data it describes, ideally in the same file. (5) Keep multiple copies in separate locations. (6) Regularly convert working files from proprietary databases to archival format. (7) Keep printed copies of critical data, with Centinel-like guard symbols and documentation for future recovery.

For full details and utility programs supporting this project, see www.cbs.umn.edu/centinel.

6. Acknowledgements

We thank Eville Gorham, Jan Janssens, Todd Lehman, Eric Lind, Richard McGehee, David Tilman, Richard Barnes, and all others who lent help and encouragement during this ongoing project. This project was supported in part by a National Science Foundation LTER grant to David Tilman and by a University of Minnesota database grant to Eville Gorham.

References

[1] L. A. Schulte and D. J. Mladenoff, "The original US public land survey records. their use and limitations in reconstructing presettlement vegetation," *Journal of Forestry*, vol. 99, pp. 5–10, 2001.

[2] J. Rothenberg, "Ensuring the longevity of digital documents," *Scientific American*, vol. 272, pp. 42–47, 1995.

[3] A. Waugh, R. Wilkinson, B. Hills, and J. Dell'oro, "Preserving digital information forever," *Proceedings of the Fifth ACM Conference on Digital Libraries*, pp. 175–184, 2000.

[4] D. Butler, "The future of electronic scientific literature," *Nature*, vol. 413, pp. 1–3, 2001.

[5] C. Tristam, "Data extinction," *Technology Review*, vol. 105, pp. 37–42, 2002.

[6] K.-H. Lee, O. Slattery, T. Lu, R. McCrary, and Victor, "The state of the art and practice in digital preservation," *Journal of Research of the National Institute of Standards and Technology*, vol. 107, pp. 93–106, 2002.

[7] S. Ong, "Worm storage is not enough," *IBM Systems Journal*, vol. 46, pp. 363–369, 2007.

[8] U. Duerig, "High density multi-level recording for archival data preservation," *Applied Physics Letters*, vol. 99, p. 023110, 2011.

[9] J. Marberg, "Towards SIRF: Self-contained information retention format," *Proceedings of the Annual International Systems and Storage Conference, Haifa, Israel*, 2011.

[10] E. T. Borer, E. W. Seabloom, M. B. Jones, and M. Schildhauer, "Some simple guidelines for effective data management," *Bulletin of the Ecological Society of America*, vol. 90, pp. 205–214, 2009.

[11] M. C. Whitlock, "Data archiving in ecology and evolution: Best practices," *Trends in Ecology and Evolution*, vol. 26, pp. 61–65, 2011.

[12] S. Rabinovici-Cohen, M. E. Factor, D. Naor, L. Ramati, P. Reshef, S. Ronen, J. Satran, and D. L. Giaretta, "Preservation datastores: New storage paradigm for preservation environments," *IBM Journal of Research and Development*, vol. 52, pp. 389–399, 2008.

[13] H. Heslop, S. Davis, and A. Wilson, "An approach to the preservation of digital records," *National Archives of Australia, Link at http:// www. naa. gov. au/ recordkeeping/ er/ digital_ preservation/ summary. html or http:// www. naa. gov. au*, 2000.

[14] R. W. Hamming, "Error detecting and error correcting codes," *The Bell System Technical Journal*, vol. 26, pp. 147–160, 1950.

[15] B. Kernighan and D. Ritchie, "The C programming language," *PrenticeHall, Englewood Cliffs, NJ*, 1978.

7. Appendix: The Centinel algorithm

The complete algorithm that encapsulates Centinel files is given here in a subset K&R C [15]. The material below, together with Kernighan and Ritchie's book, should allow the algorithm to be transcribed into future programming languages and the data to be extracted from Centinel files as long as the printed form is extant.

The algorithm adds an error-correcting code to each line of a text-based file, another to each page, and a third to the entire file. Each output line begins with a decimal error correcting code guarding that line, and also guarding the error correcting code itself, then the text of the line. In printed form another decimal code guards the entire page and a third guards the entire file.

In computing the error correcting code, leading and trailing white space is skipped, multiple blanks count as a single blank, and end-of-line codes are not counted. The code at the beginning of the line is not counted either. The assignment between symbols and numbers is specified in array s below, where 'a' is number 1, 'b' is number 2, 'A' is number 27, and so forth. Any similar assignment could be substituted.

In the algorithms below, flow control and reserved words are bolded, variables and function names are italicized, and certain operations such as '<=', '>=', '!=', and '==' are displayed in a mathematical form as '$\leq$', '$\geq$', '$\neq$', and '$\equiv$', respectively.

DATA STRUCTURES

```
#define C      256          1. Maximum character code plus 1.
#define L      120          2. Maximum data length, excluding guard symbols.
#define G      8            3. Number of guard symbols.
#define COL    9            4. Number of symbols columns displayed on the page.
#define PAGEL  50           5. Number of lines per page.
#define IDENT  127          6. Identity symbol.

char s[] =                  7. Character set available for present application.
  "_abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMNOPQRSTUVWXYZ0123456789"
  " .,;:!?+-*/\\=\"' ()[]{}<>^&%|";

int nchar;                  8. Maximum number of characters in present application.
char seq[C];                9. Sequence number for each symbol in the set.
char f[C][C];               10. Modulo sum and difference tables.
char ptn[L][G + 1];         11. Pattern of guard symbols for each position.
int pagef = PAGEL;          12. Number of lines on first page.
int pages = PAGEL;          13. Number of lines per subsequent page.

int ipage = 0;              14. Page index.
int ifile = 0;              15. File index.
char in[L + 1];             16. Input line.

char line[L + 1], page[L + 1], file[L + 1];   17. Current line, page, and file.
char guard[G + 1];          18. Guard symbols, individual characters.
```

END OF PAGE

Upon entry to the algorithm, (1) $page$ contains a list of symbols representing the current page. **(2)** $ipage$ indexes the next entry for the page. **(3)** a is set if a blank line should follow the code, indicating end of page. (This is not used on the last page of the file, because the code for the entire file follows immediately.) **At exit, (1)** Guard symbols for the page are displayed. **(2)** $guard$ is destroyed. **(3)** $ipage$ is set to zero.

```
seqpage(a) int a;
{
  if (ipage ≡ 0) return;
  page[ipage] = 0; ecc(guard, page);
  seqn(guard, "; ", ""); if (a) printf("\n");
  ipage = 0; }
```

MAIN PROGRAM

```
main(argc, argv) int argc; char *argv[];
{ char c; int i, j, k;

  if (argc > 1)
  { pagef = atoi(argv[1]);
    if (pagef < 2 || pagef > 100) pagef = PAGEL;
    pages = pagef; }

  if (argc > 2)
  { pagef = atoi(argv[2]);
    if (pagef < 2 || pagef > 100) pagef = PAGEL; }

  s[0] = IDENT;
  for (i = 0; s[i]; i = i + 1) seq[s[i]] = i;
  nchar = i;

  for (i = 0; i < C; i = i + 1)
  for (j = 0; j < C; j = j + 1)
    f[i][j] = IDENT;

  for (i = 0; s[i]; i = i + 1)
  for (j = 0; s[j]; j = j + 1)
  { k = i + j; if (k >= nchar) k = k - nchar;
    f[s[i]][s[j]] = s[k]; }

  for (i = 3; i <= 7; i = i + 2) colgen(i, G - 1);
  ipage = 0; ifile = 0;

  while (fgets(in, L, stdin))
  { i = strlen(in);
    if (in[i - 1] == '\n') in[i - 1] = 0;

    line[0] = '-';
    for (i = j = 0; in[i]; i++)
    { c = in[i]; if (seq[c] == 0) c = ' ';
      if (line[j] == ' ' && c == ' ') continue;
      line[++j] = c; }
    line[++j] = 0;

    ecc(guard, line + 1); seqn(guard, "", in);

    page[ipage++] = guard[G - 1];
    if (ipage >= pagef) seqpage(1), pagef = pages;

    file[ifile++] = guard[G - 1];
    if (ifile >= L) ifile = ifile - 1; }

  seqpage(0); file[ifile] = 0;
  ecc(guard, file); seqn(guard, ".", "");
  ifile = 0; }
```

1. If an entry parameter has been supplied, take it to be the page length.
2. Determine the number of symbols in the set while developing a list of sequence numbers.
3. Clear the modulo addition table.
4. Construct tables mapping all symbol pairs to corresponding sums.
5. Generate odd guard patterns.
6. Compute the error-correcting code for the line.
7. Compress multiple blanks from the input line.
8. Compute the ECC guard symbols.
9. If this is the end of the page, prepare a code for the entire
10. If this is the end of the page, prepare a code for the entire
11. At the end of the file, prepare a code for the entire file.

COMPUTE CENTINELS

Upon entry to the algorithm, (1) gs points to an area of length $G + 1$ to receive the results. **(2)** $line$ points to the line. **(3)** G defines the number of guard digits to be computed. **(4)** ptn defines which line positions contribute to which guard digits. **(5)** f contains the modulo-addition table for all symbols. **At exit,** gs contains the guard symbols for the line.

```
ecc(gs, line) char *gs, *line;
{ int i, j;

  for (i = 0; i < G; i = i + 1) gs[i] = IDENT;

  for (i = 0; i < G; i = i + 1)
  for (j = 0; line[j]; j = j + 1)
    if (ptn[j][i] == 'X')
      gs[i] = f[gs[i]][line[j]];
  gs[G] = 0; }
```

1. Clear all the guard symbols.
2. Generate each guard symbols.
3. following the table that shows which line positions contribute to which guard symbols.

CONVERT CENTINELS TO INTEGERS

Upon entry to the algorithm, (1) gs contains the guard symbols. **(2)** sep contains a separator character. **(3)** sym contains the string of symbols. **At exit,** gn contains the corresponding integer sequence numbers.

```
seqn(gs, sep, sym) char *gs, *sep, *sym;
{ int i;
  for (i = 0; i < G; i = i + 1)
    printf("%02d", seq[gs[i]]);
  printf("%s%s\n", sep, sym); }
```

1. Display the sequence numbers for the guard symbols.
2. Display the full line.

GENERATE PERMUTATIONS

Upon entry to the algorithm, (1) n defines the number of guard symbols to be marked. **(2)** k defines the position for the initial mark. **(3)** l defines the column number on the line, starting with 0. **(4)** ptn contains an area to receive the permutations. **(5)** w contains a work area for generating the permutations. **At exit, (1)** All permutations have been generated. **(2)** l is advanced by the number of combinations generated. **(3)** $ptn[0..l]$ contains the permutations generated thus far. **(4)** w contains the most recent permutation generated.

```
colgen(n, k) int n, k;
{ static char w[G + 1] = ""; static int l = 0; int i;
  if (w[0] ≡ 0)
    for (i = 0; i < G; i = i + 1) w[i] =' -';
  if (n > 0) for (i = k; i ≥ n - 1; i = i - 1)
  { w[i] =' X';
    colgen(n - 1, i - 1);
    w[i] =' -'; }
  else if (l < L)
  { for (i = 0; i < G; i = i + 1)
      ptn[l][i] = w[i];
    l = l + 1; }}
```

1. On the first call, establish a null pattern in the array.
2. Mark the guard symbol for each possible position and generate all permutations within that position.
3. If there are no deeper permutations, save the current permutation and advance the column number.

Study of Database Key Problems in Network Printing on Demand Modeling

A. Zhu Jia[1], **B. Li Ye-li**[2]
[1,2] Information and Mechanical Engineering Institute, Beijing Institute of Graphic Communication, Beijing, China

Abstract - *In a workflow management system, database technology is a significant component to support asynchronous and distributed collaborative working, and database also is an indispensable part in workflow management system. In the business process modeling, it is very important to exchanging information with databases. This paper mainly discussed several key problems that have to do with operation of database in constructing the online printing on demand model in Business Process Management engine, and then for each problem we found out the relevant solutions.*

Keywords: Process Modeling, Database Operation, Printing On Demand.

1 Introduction

Develop a system, while the key part of it is the workflow engine that is how to develop the program logic according to the actual need and make sure that the stability, easy maintenance and elastic. But the system interface and exchange of information with database are also very important. In workflow management system, database technology is an important component that supports asynchronous and distributed collaborative working. Workflow management system is an indispensable part of the foundation. The operation of the workflow database is to enhance the transparency of the system, which can realize the workflow process to monitor, workflow restructuring and workflow reuse target, etc. Database is the collection of related data according to certain rules of structure and organization. While the user inquires the readings on interface in the process of the establishment of online printing on demand model, the data was put in the database in advance, and finally the website will return the operating results to the browser and the user will see the answer through the browser. The data from writing in to read out, we will meet a series of problems about database security, backup, restore, filter and so on. In order to guarantee the stability of database in the modeling process, it is necessary to operation of the database correctly from the beginning. This article in view of operation some database problems about the modeling process to find out the corresponding solution and get the expected results.

2 Condition filter about MYSQL in workflow software

In the modeling process, after adding a database connector we need to design the form, which is to let the reading display on the interface in different conditions. There appeared one problem to transfer the parameters. That is to say, when doing the search page, we often meet many conditions to inquiry and these conditions are unsteady, when the condition was input by user is null, it should not be added to the SQL statement. For example, we want to search for a bibliography book, the conditions are may: A, price for 100; B, the title of the book includes key word P, then we may write condition as WHERE price = 100 AND name LIKE '%P%', the problem here is the price 100 and key word P are selected or input by user. When the user did not have a choice or enter a condition among them, the filter conditions also shouldn't be there, so we will need logical judgment in the pages. The more conditions we use, the more IF statements will appear, and the page will appear a lot of logic SQL statements to combine. It obviously increased the workload and the difficulty about maintenance the programming code.

One solution to this problem is: this kind of complex SQL statements were written in a special SQL template. In view of the above mentioned problems, the related SQL statement template should write as follows: SELECT * FROM book WHERE price = IF('{0}' = '', price, '{0}') AND name LIKE IF('{1}' = '', name, '%{1}%'); It means if the values {0} is null, price=price, if it is not null, price= {0}. For the name, it is also the same meaning. The price and name were two fields in the products table, {} flag is the place we will replace the parameters. So that the inquiry page just call SQL template statements and replace the corresponding parameters. If user didn't want to filter price, the value {0} is empty, so we get the SQL statement similar to: SELECT * FROM book WHERE price = price AND name LIKE 'P', among them, 'price=price' didn't have the role of the filter. So that we reached the purpose of dynamic generating more conditions to inquire and the pages wouldn't need to have the combination of more SQL statements.

3 Use the RowSet interface to visit data sources

This RowSet object can establish a connection with data source and maintain the connection throughout its life cycle. In this case, the object is called as the connection RowSet. Rowset can also set up a connection with data sources to get the data and then close it, this rowset are called as the connectionless rowset. The connectionless rowset can change its data when disconnected, and then these changes will send back to the original data sources, but it must be to establish a connection to complete this operation. Compared to the Java.sql.ResultSet, offline operation of RowSet can effectively use computer's plenty of memory and reduce the burden of the database server. Because the data operations are in memory and then submit them to the data source, the flexibility and the performance have improved a lot. RowSet default can scroll and update and serialize result sets, and it as JavaBeans can be easily transmission in the network and used for data synchronization between both ends. The exact understanding is that we can look at the RowSet as object has nothing to do with the database but only represent by data, which is related to a problem where the data come from. Since most of the cases that dealing with the data is equal to dealing with database, so RowSet interface provided a method access data directly from database through JDBC.

In printing on demand process modeling, we operate the database to hope that the inside data of the database can display on the interface one line by one line. Due to the flow software was compiled by Java code, this project use rowSet interface to operate data in database. For example, rowSet.tolist() is equivalent to take out a column information of many lines inside of the database, and then can be placed into a variable; rowSet.getValues() means to show a whole record that we inquire from the database.

4 The Security and Timeliness of database

Database security means to protect the database from illegal use which will cause data leak, changes, or destruction. User name marking and identification is on the one hand, to regularly back up the database is on the other hand. Once the database destroyed, we can through the backup file to restore database. Back up the database have many ways and the effect is not the same. It needs to use the backup data if the database appeared faults, which let the loss to lowest. In the network printing on demand process modeling, we add the MYSQL database connectors in order to query the existing bibliography in the database. When a bibliography and material database have built, it involves the question either the database into or out while migrate it from one system to another. If we encounter creating a new database, the trouble is to enter a lot of table and records, so we can use existing database to import, which save a lot of time and also avoid error.

The mysqldump command can backup the database into a text file. The table structure and the table data will be stored in the text files. The principle of this command is very simple. First it find out the structure of the table that need to backup, again generates a CREATE statement in the text file. And then all of the records in the table will be converted into an INSERT statement. These CREATE statements and INSERT statement is used to restore. The basic grammar of using mysqldump command to backup a database is: mysqldump -u username -p dbname table1 table2...>BackupName.sql; Among them, the dbname parameter represent the name of the database; table1 and table2 parameters represent the name of table, all the database will backup if not have the parameter; BackupName.sql parameters represent the name of the backup file, before the file name we can add an absolute path which the database will backup into a sql suffix file name. For example: the command that with the user of root backup the student table in the test database is: mysql -u root -p test student > e:\student.sql; After command you will find the student.sql file at E dish.

The command of backup multiple databases is: mysqldump -u username -p --databases dbname1 dbname2...>BackupName.sql；Among them, dbname1 and dbname2 represent the name of database, it will backup all the databases if use the all databases parameter. For example: the command that with root user backup test database and mysql database is: mysqldump -u root -p --databases test mysql > c:\backup.sql. After command, we can find the backup.sql file in the C dish, all the information of these two databases are in this file.

When migration the database from one system to another, it will appear 3 kinds of situation, they are: migration in the same version of the mysql database, move to other versions of the mysql database and migrated to other types of database. Of course, the easiest way is moving between the same version of the mysql database. In order to keep the same version, we use the same system, and then use the following command reduction the database before: mysql -u root -p [dbname]< backup.sql; For example, use the mysql command to reduction the score table in test database: mysql -u root -p test < d:\backup\score.sql; The score.sql file must on the D dish if we use this command. Note that all the above operations are input by the cmd format.

5 The solution about messy code problem of database

When operating all those above steps, we may inquire the current databases, the most likely problem is messy code. This kind of problem is the questions about MYSQL character set support. MYSQL Character Set Support has

two aspects: Character Set and Collation. Specific to the following four levels: server, database, table, connection.

5.1 MYSQL default character set

MySQL character set can be refined to a database, a table, and a column. Traditional program didn't use so complex configuration to build database and data table.It use the default configuration. When compile MySQL, it appointed a default character set that is the latin1; when install MySQL, it can be assigned default character set in the configuration file my.ini. If not specified, this value will inherit from compiling; When start MYSQL, it can be specified one default character set in command line parameter, if not specified, this value will inherited from the configuration in its configuration files, and right now character_set_server is set to the default character set; When creating a new database, the character set of this database was defaulted for character_set_server unless specifically designated; When selected a database, character_set_database was set for the default character set of this database; To create a table in this database, the default character set of this table is set to character_set_database, also it is the default character set of database; When set one bar in table, the default character set of this column is the default character set of this table unless specifically designated; If we do not modify any place, then all column of all the table among all the databases will be stored with latin1, but if we install MySQL, general we will choose more language to support, that is to say, the installation program will automatically set the default_character_set to utf-8 in configuration file, this ensures that all column of all the table among all the databases use utf-8 storage in the default situation.

5.2 Check the default character set

In default situation, the character set of MySQL is latin1, check the character set of system and sorting order can be usually set through the following two commands:

(1) mysql> SHOW VARIABLES LIKE 'character%'; The results are shown in Fig.1.

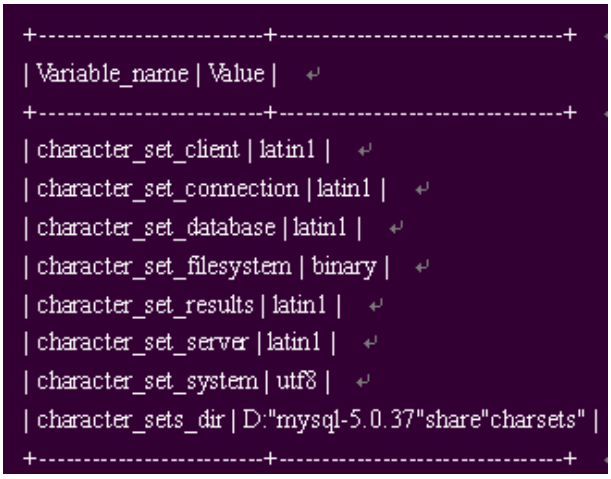


Figure 1 check the database character set

(2) mysql> SHOW VARIABLES LIKE 'collation_%'; The results are shown in Fig.2.

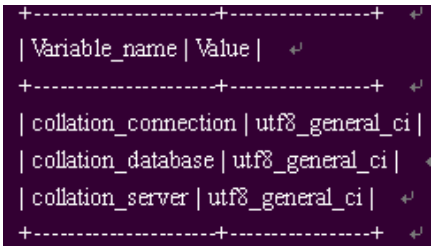


Figure 2 system sort order

5.3 Modify the default character set

The most simple method is to directly modify the key value of character set in my.ini file, such as default-character-set = utf8 ， character_set_server = utf8. After modification, restart MySQL service, use the command as mysql> SHOW VARIABLES LIKE 'character%'; We will find out that the database coding has been turned into utf8 as the results in Fig.3.

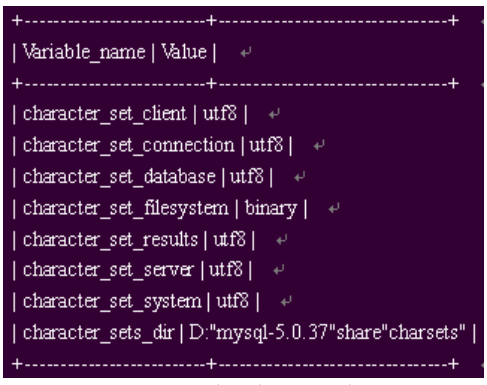


Figure 3 database coding

Still we have a more trouble method to get this result that is to use MySQL command:

SET character_set_client=utf8;

```
SET character_set_connection=utf8;
SET character_set_database=utf8;
SET character_set_results=utf8;
SET character_set_server=utf8;
SET collation_connection=utf8;
SET collation_database=utf8;
SET collation_server=utf8;
```

Generally even if we set the default character set of table for utf8 and send query through the utf-8 coding, you will find the database was still garbled. The main problem is the connection link layer. The solution is to execute a sentence before inquires: SET NAMES 'utf8'; It is equivalent to the following three instructions:

```
SET character_set_client=utf8;
SET character_set_results=utf8;
SET character_set_connection = utf8;
```

6 Conclusion

Workflow is a formalization description of business process, including describes the workflow logic of the dependent relationships among tasks and increases the workflow semantics of dominant content on this logic. The database has many strong functions such as data organization, user management, the security check and so on. The printing on demand flow model can't realize without combining the database technology and workflow technique. If we can accurately and masterly operation the database, it will be lay a solid foundation for the future printing on demand system's realization.

Acknowledgment

The authors would like to thank "Funding Project for Academic Human Resources Development in Institutions of Highter Learning Under the Jurisdiction of Beijing Municipality. （PXM2010_014223_095557）".

7 References

[1] Workflow engine on http://baike.baidu.com/view/1636259.htm.2011-11-1.

[2] Christopher Keene, in: Migrating From MS Access to MySQL, 2007(8).

[3] Zhao Li-xiang, Yin Guo-fu and Shu Bin, in: Flexible Workflow Modeling Based on Database Management. Computer Integrated Manufacturing Systems[J]. 2003, 9(2).

[4] Huang Jin-hua, in: Mysql introductions are so easy[M], Tsinghua university press,2011(1), p.352-370

[5] Li Xu-dong, Cheng Ren-hong, Tu Feng-sheng, in: Access And Migration Among Databases Based on Different Charcters[J], Computer applications, 2001,12,(12), p.33-36.

[6] Bonitasoft | Open source Business Process Management (BPM) and Workflow software on http://www.bonitasoft.com/，2011-6-24.

Processing Continuous Range Queries on Moving Objects using a Bit Vector Based Index*

I-Fang Su[1], Chang-Ming Tsai[2], Yuan-Ko Huang[3], Chieh-Ling Huang[2], Jwu-Jenq Chen[2], and Yu-Chi Chung[2, a]

[1]Department of Information Management ,Fortune Institute of Technology, No.1-10, Nongchang Rd., Daliao Township, Kaohsiung County 831, Taiwan (R.O.C.)

[2] Department of Computer Science and Information Engineering, Chang Jung Christian University, 396 Chang Jung Rd., Sec.1, Kway JenTainan 71101, Taiwan (R.O.C.)

[3]Department of Information Communication, Kao Yuan University, No. 1821, Jhongshon Rd., Lujhu, Dist, Kaohsiung City 82151, Taiwan (R.O.C.)

Abstract - *Handling continuous range queries over moving objects has caught wide attention recently. As a set of continuous range queries are submitted, the server periodically evaluates these queries and reports objects that are currently within the query boundaries. We propose a new index structure, called Continuous Range Queries Indexing based on Bit Vector (CRQIBV), for processing continuous range queries over moving objects. Performance evaluations show that CRQIBV can achieve better performance than other index structures (e.g., CES).*

Keywords: Data structures, query processing, query index, continuous range queries, moving objects.

1 Introduction

With the advances in technology, many wireless devices such as cell phone, tablet computer and notebook computer are equipped with location-sensing components. Therefore, location-aware applications are booming. These applications are designed to monitor user location and provide related services to the users. For example, businesses can dispatch discount coupon to users in the vicinity of their stores. A cab operator can obtain the location of a passenger and quickly dispatch a nearby taxi to the passenger. A teacher can use hand-held device to monitor the location of a group of children in a museum area.

In order to make services available to the users in such applications, we need an efficient mechanism of continuous range query processing. Given a query region, a continuous query will keep track of the moving objects in the region. In the example of dispatching discount coupon, a business operator can designate an area around the store as its query region and acquire information of moving objects (i.e., the shoppers) that enters this region.

To process continuous range queries efficiently, we are expected to confront with several issues. First, due to the fact that objects may move, applications need to reevaluate queries periodically to ensure the correctness of query results. Second, a system may execute many continuous queries concurrently. Therefore, the system must attain a very short execution time for large number of queries.

There are many studies regarding the continuous range query processing in the past [1,2, 4-11]. All of them use indexing techniques for supporting efficient processing of continuous range queries. These indexes can be classified into two categories. One is object indexing [7-11] which builds an index on the moving objects. The other one is query indexing[1,2, 4-6] which builds and index on the query objects. It has been proved that query indexing is a more efficient approach for continuous range query processing[1,2]. The main reason lies in the fact that there are more moving objects than query objects in this type of applications. Therefore, compared with the query indexing, object indexing will have higher maintenance cost. As a result, the overall system efficiency is compromised.

Kalashnikov *et al.* [2] proposed a cell-based query index structure. It partitions the whole monitoring region into cells. Each cell is associated with two different query lists, *full* and *partial*. When a query is placed in the full query list, the range of this query covers the entire cell. When a query is placed in the partial query list, the range of this query covers only part of the cell. We use Figure 1 to clarify the meaning of these terms.

In Figure 1, the entire monitoring region is divided into 5x5 cells. There are two queries in this monitor region. Using cell(3, 3) as an example, its full query list and partial query list are shown in Figure 1. The full query list of cell(3, 3) contains query Q_1. This is because the query region of Q_1

[a] Corresponding author.

* This work is supported by National Science Council of Taiwan (R.O.C.) under Grants NSC 100-2221-E-268 -007, NSC100-2221-E-309-011, NSC 100-2221-E-244 -018.

fully covers cell(3, 3). On the other hand, Q_2 only covers part of cell(3, 3), so that Q_2 is placed in the partial query list of cell(3, 3).

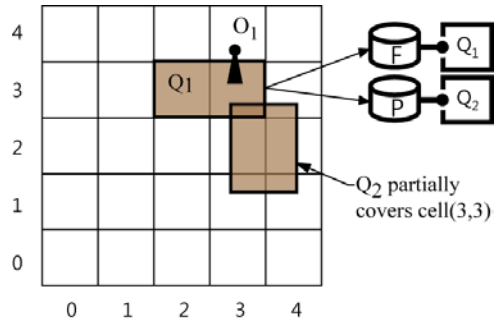


Figure 1. An example of the cell-based query index.

Assume that object O_1 appears in cell(3, 3), the algorithm first retrieve query lists of cell(3, 3). It is found that Q_1 is in the full query list. Therefore, we know that Q_1 covers cell(3, 3) completely. Hence, object O_1 must be part of result of query Q_1. Since Q_2 is in the partial query list, meaning that Q_2 only covers part of cell(3, 3). Therefore, we need to extract the query region of Q_2 and compare the region with O_1's location in order to identify that Q_2 truly contains O_1. The above process requires intensive computation, which makes frequent query reevaluations more difficult to conduct.

K. L. Wu *et al.* propose several improved query indexes such as shingle-based index [6], covering tile-based index [5], VCR-based index [4] and CES-based index [1]. Since the experiment results of [1] shows that CES-based index performs better than the other three indexes. We shall only explain the details about CES-based index in the following.

CES stands for *containment-encoded squares*. CES decomposes the monitor region into cells. For each cell, CES further splits each cell in a quad-tree manner (Figure 2). Every node of the quad-tree is a *virtual construct rectangle* (*VCR*). CES requires each side of a leaf node is one unit in length (i.e., a 1 × 1 node). According to the partition rule of a quad-tree, a level-*i* node is a VCR of size $L/2^i \times L/2^i$, where L is the side length of the root node (i.e., the level-0 node). Let k be the number of levels in an quad-tree, then we can obtain $k = \log_2(L)$. For example, in Figure 2, $L = 4$ and $k = \log_2(4) = 2$. The root node (i.e., the node at level-0) is a VCR of size 4×4. Every node in level-1 is a VCR of size 2×2, and every level-2 node is of size 1 × 1. Given a query, CES decomposes its query region into several VCRs. As an example, Figure 3 shows that the query region of query Q_1 is decomposed into three VCRs: one 4×4 VCR and two 2×2 VCRs. The query ID (i.e., 1) is then inserted into the query ID lists associated with the decomposed VCRs.

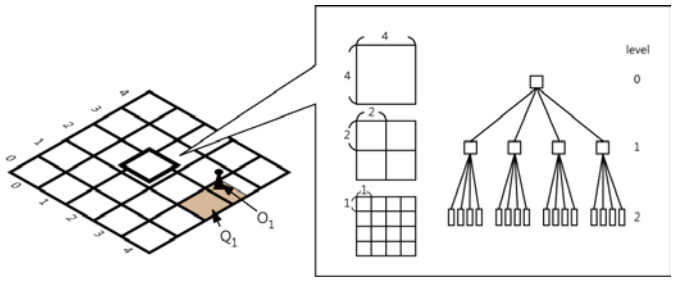


Figure 2 CES-based index.

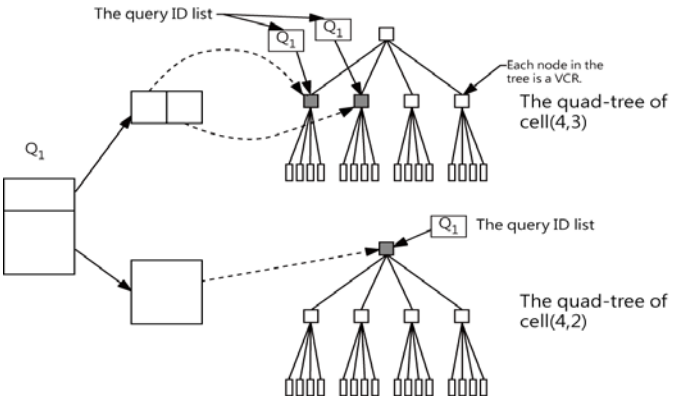


Figure 3 Q_1 is decomposed into three VCRs.

Given an object location (x, y), CES search algorithm operates as follows. First, the VCRs that contain (x, y) are identified. Then the IDs inside these VCRs' query ID lists are retrieved. At this point, the set of queries that contains (x, y) is determined. Again, we use an example to illustrate the process. In this example, the location of object O_1 is shown as in Figure 4. The cell that O_1 resides can be quickly identified through the location of O_1. In this case, it is cell(4, 3). The CES search algorithm then performs a bottom-up manner to visit the quad-tree in cell(4, 3). We show the visiting order in Figure 4 (i.e., CES first visits V_1, then V_2, finally, V_3). For a VCR that is visited during the search, the IDs in its query ID list are extracted. Finally, we can determine the covering queries for O_1. In this example, we know that Q_1 is a covering query for O_1. Hence, O_1 becomes the result of query Q_1.

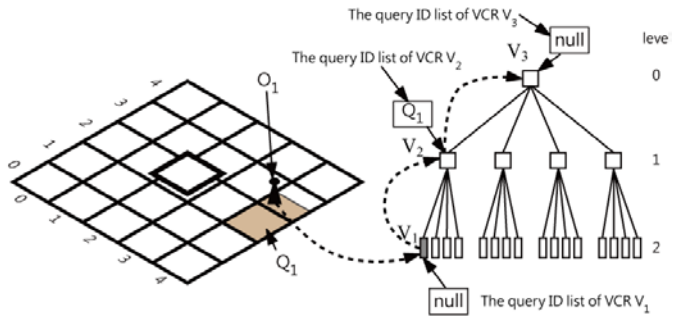


Figure 4 A query evaluation example of CES.

The drawback of CES is that the query decomposition algorithm needs intensive computation. When executing query operation such as insert and delete, it adds significant load to the database system. In addition, when the number of queries grows, the number of VCRs quickly increases accordingly. The whole CES index becomes too big to be fit into main memory. As a consequence, additional I/O costs are incurred. Finally, given an object, whether it is covered by any query or not, the CES search algorithm always traverse the entire quad-tree in a bottom-up manner. However, we found that most of the objects in the monitoring region have not been covered by any query. For these object CES takes $O(k)$ time complexity in order to determine if they are not covered by any query, this wastes a lot of computing resources.

In this paper, we design a query index structure to make the processing of continuous range queries efficient and scalable. We use bit vectors as the underlying data structure of the proposed index. Bit vectors take very small amount of memory, therefore they are memory efficient. Meanwhile, the bitwise operations are very fast to execute and can often be accelerated by hardware [3]. Besides, our design of index can quickly identify those moving objects that are not covered by any query. This further improves system's performance. In the experiment, we compare the performance of our query index and CES. The results show that our query index structure outperforms CES in terms of the computation cost, and the storage cost.

2 Preliminary

In this section we introduce some notation for the study of the query indexing mechanism and state the assumptions made in this paper. In this paper, Q is a set of continuous range queries and O is a set of moving objects. We use q to denote one of the continuous range queries in Q. We use the notation (l_0, l_1) and (u_0, u_1) to represent the coordinates of bottom-left and top-right corners of q, where l_0, l_1, u_0, and u_1 are integers. In our method, every query submitted to the server is assigned an identity. We use q_k to represent the query whose id is k. Similar to [1, 2], we assume that the continuous range queries are stationary but they can be inserted or deleted dynamically. Objects move continuously and no constraints are imposed on the speed or path of moving objects [2]. Let M be the monitoring region where moving objects are tracked. Without loss of generality, we assume M is the [0, L] x [0, L] square on the coordinate plane for the following discussion.

2.1 Data structures

In this paper, we proposed a method called *Continuous Range Queries Indexing Based on Bit Vectors* (*CRQIBV*). This method employs bit vectors as the underlying data structure. Due to the compact sizes of bit vectors and relatively fast bitwise operations, the CRQIBV method should be more efficient in both memory usage and computation needs. Figure 5 shows the data structures used in CRQIBV.

One of them is the *Query Buffer* (*QB*). It is a hash table that caches all active continuous range queries, with query id serving as its key attribute. In Figure 5, there are two queries, q_0 and q_1, running on the server. As a result, *QB* stores the context of these two queries. *Query ID pool* (*IDPool*) is a container that stores the query identities that CRQIBV can dispatch. In Figure 5, *IDPool* only has one identity remaining (i.e., 2). When a new query q is submitted to the system, CRQIBV assigns 2 to the identity of q. CRQIBV logically divides the monitoring region [0, L] x [0, L] into L equal *stripes* along x-axis and y-axis respectively. The j-th vertical stripe is denoted as $D_0.S_j$, and the j-th horizontal stripe is denoted as $D_1.S_j$,. j = 0, 1, 2, …, L-1. We create a bit vector $BV(D_i.S_j)$ for the stripe $D_i.S_j$ (i = 0, 1 and j = 0, 1, 2, …, L-1) when some query region overlaps with the stripe. If the query region of query q_k intersects the stripe $D_i.S_j$, then the k-th bit of $BV(D_i.S_j)$ is set to 1, otherwise, the bit is set to 0.

As shown in Figure 5, a number of stripes do not have a bit vector (e.g., $D_0.S_3$, $D_0.S_4$, $D_1.S_0$, $D_1.S_1$, and $D_1.S_2$) created for them. This is because that there are no overlaying query regions for these stripes at this moment. If a monitoring region is large, then numerous stripes may not have any overlapping query region and, therefore, have no bit vectors associated with them. This bit vector allocation strategy can thus save substantial memory space. The length of bit vectors discussed above is a system parameter of CRQIBV, and is denoted as *BL*. It specifies the maximum number of queries that can be traced in the bit vectors. In Figure 5, *BL* = 3.

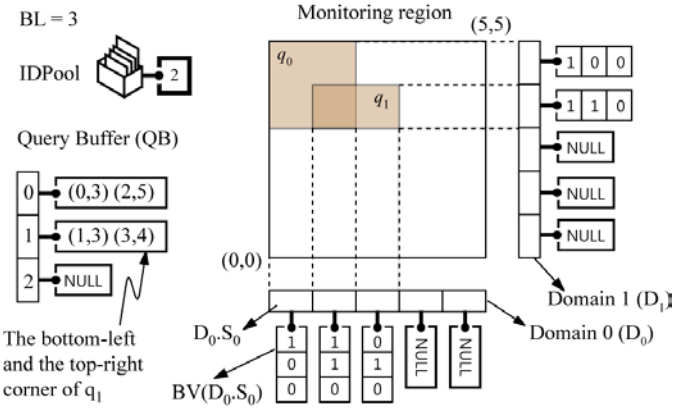


Figure 5 The datastructures used in CRQIBV.

3 CRQIBV

In this section, we explain how CRQIBV performs query insertion, query deletion, and query evaluation. Detailed algorithm is omitted due to space limitation. Instead, we explain the algorithm using an example.

Query insertion. When CRQIBV receives the query q submitted by a user, (1) it retrieves an identity (say, k) from the IDPool and assigns it to q; (2) it inserts q_k into the query buffer; and (3) it searches every stripe for the intersection with

the query region of q_k and sets the k-th bit in the bit vector accordingly. In Figure 6, CRQIBV receives a new query q. First, CRQIBV searches the IDPool for the smallest identity and assigns it to q. In this example, an id of 2 is assigned to q and we call this new query q_2. Then, "2" is deleted from the IDPool. After this, IDPool becomes empty. Then, CRQIBV inserts this new query into QB using 2 (i.e., the id of q_2) as its key value.

We assume that the bottom-left corner (i.e., (l_0, l_1)) and the top-right corner (i.e., (u_0, u_1)) of q_2 are (3,1) and (4,2), respectively. CRQIBV uses (l_0, l_1) and (u_0, u_1) to determine which strip intersects with the query region of q_2. For D_i, its strips $D_i.S_{li}$ to $D_i.S_{ui-1}$ intersect with the query region of q_2. In the case of D_0, $D_0.S_{li}$= $D_0.S_3$ and $D_0.S_{ui-1}$= $D_0.S_{4-1}$= $D_0.S_3$; therefore, CRQIBV sets the second bit of $D_0.S_3$ to 1. Because no query intersects with $D_0.S_3$ before q_2 appears, CRQIBV allocates a bit vector for $D_0.S_3$ in advance and then set the second bit of BV($D_0.S_3$) to 1. The result is shown in Figure 6. Similarly, CRQIBV also allocates a bit vector for $D_1.S_1$ and sets the second bit of BV($D_1.S_1$) to 1.

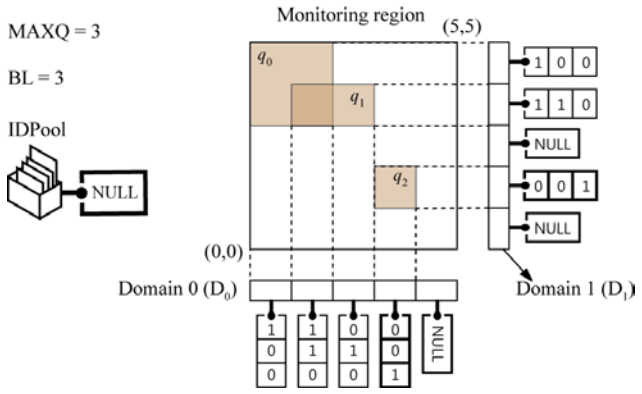


Figure 6 q_2 is inserted into the CRQIBV.

Query Deletion. After query q_k expires. CRQIBV executes the following steps to remove q_k from the index structure: (1) put the identity of q_k back into the *IDPool*; (2) remove q_k from the query buffer; and (3) find all stripes that intersect with the query region of q_k and clear the k-th bit of their bit vector to 0. In Figure 7, assuming q_1 has expired, CRQIBV returns 1 to the IDPool and remove q_1 from QB. Then CRQIBV scans all strips that intersect with the query region of q_1 and sets the first bit of the bit vector to 0. In this case, the first bit of $BV(D_0.S_1)$, $BV(D_0.S_2)$, and $BV(D_1.S_3)$ is set to 0. Additionally, CRQIBV discovers that every bit of $BV(D_0.S_2)$ is 0; thus, it releases the memory occupied by $BV(D_0.S_2)$ to minimize the memory consumption of CRQIBV.

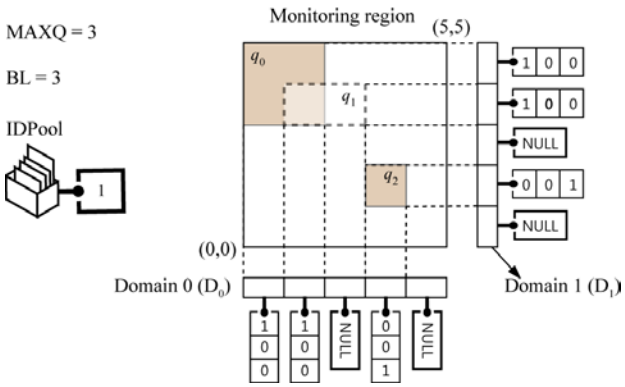


Figure 7 q_1 is removed from CRQIBV.

Query evaluation. For a given object, we use Figure 8 to explain the use of CRQIBV for checking which queries cover this object. In this example, there are two moving objects o_1 and o_2 in the monitoring region. Their locations are shown in Figure 8. We use $Loc(o)$ to represent the location of the moving object o. In Figure 8, $Loc(o_1) = (p_0, p_1) = (1.5, 3.5)$. Here, p_i represents the i-th coordinate value of o_1. Along the x- or y-axis, we can identify the stripe that contains o_1 through the floor function floor(p_i/SL), where SL is the width of stripes. In our example, SL = 1 and floor(p_0/SL) = floor(1.5/1) = 1. This implies that $D_0.S_1$ contains o_1. Similarly, we can determine that o_1 is located in $D_1.S_3$ too. After performing bitwise AND operation on BV($D_0.S_1$) and BV($D_1.S_3$), we can identify which queries cover o_1. In this case, BV($D_0.S_1$) & BV($D_1.S_3$) = 110 & 110 = 110. The 0th bit and the 1st bit are 1; therefore, q_0 and q_1 cover o_1. Using the same method, we discover that o_2 is located in $D_0.S_4$ and $D_1.S_3$. But, BV($D_0.S_4$) = NULL, this implies that no query region intersects with $D_0.S_4$; thus, no query covers o_2. The o_2 example shows that CRQIBV can quickly determine if a certain object is covered by a query or not. If a monitoring region is large, then the queries are likely to be distributed sparsely and many moving objects may not be covered by any query. This mechanism can improve the efficiency of query evaluation of CRQIBV in such scenarios.

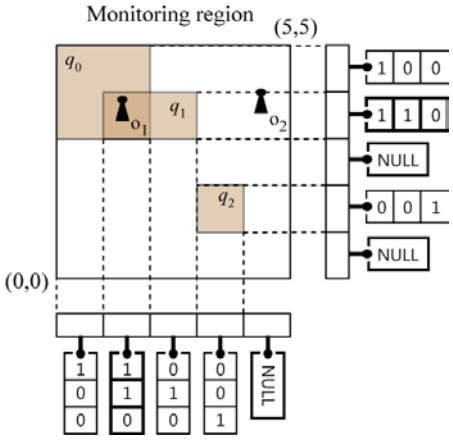


Figure 8 CRQIBV query evaluation.

The current version of CRQIBV cannot deal with the case that the number of active range queries is larger than *BL*. In previous example, *BL* = 3, meaning that CRQIBV can only index at most three queries. One solution is to set *BL* as large as possible so that there is always enough memory space to index all incoming queries. However, this method may waste memory when the number of queries is much smaller than *BL*. A better solution is to construct and extend the bit vector dynamically. We intend to explore this issue further in future work.

4 Performance

We conduct simulations to evaluate the effectiveness of the CRQIBV and to compare it with CES. Similar to [1], the monitoring region is defined as a 512 by 512 grid. All simulations run for a period of 1000 seconds.

Our simulator constructs a continuous range query q based on five parameters: $<W, H, (x, y), t_e, t_l>$. W and H are the width and the height of the query respectively. Our simulator selects W and H uniformly at random in [1, QS]. QW (stands for the query window size) can vary from 40 to 160 with a default value of 80. The coordinates (x, y) is the bottom-left corner of q. In the simulation, all queries are uniformly distributed in the monitoring region. That is, we choose x and y, uniformly and randomly, in [1, 512]. t_e denotes the entry point in time for q, which indicates the time that the user submits this query. t_e is an integer which is randomly selected from 1 to 1000. t_l is the life time of q. Therefore, a query is active during $[t_e, t_e+t_l]$, and it should be dropped from the index at the time t_e+t_l.

A total of $|O|$ objects are generated. Like continuous range queries, the initial locations of all objects are uniformly distributed in the monitoring region. We adopt the moving pattern in [1] to decide the subsequent locations of moving objects. We omit the details due to the space limitation, interested readers can refer to [1] for more detail.

In CES, the cell size is an important parameter which affects the storage cost and total evaluation time of the query index. In our simulation, we set cell size to 16 as the simulations of [1] suggest. It has been shown that CES can achieve the best evaluation time without incurring a substantial storage cost when the cell size is 16. In CRQIBV, we set *BL* = 125, which means that each bit vector has 125 bits. We choose *BL* = 125 because the number is long enough for CRQIBV to index all queries and will not incur overflows.

The simulation proceeds as follows: First, queries and objects are generated and put into buffers. The index (CES or CRQIBV) is then initialized and all objects are added into the monitoring region. At each cycle, we perform the following five steps to evaluate query results. (1) The simulator updates objects locations. (2) We scan the query buffer and retrieve all queries whose entry points (i.e., t_e) are equal to the current time. (3) The queries obtained from the previous step are inserted into the index. (4) Remove all expired queries from the index. (5) We perform query evaluation to obtain query results. We measure the average time needed to complete the evaluation of query indexes, the average update time of the indexes, and the total storage cost for the query indexes. All the algorithms are implemented in Java and the experiments are performed on a Windows Vista system with an Intel Core 2 CPU (2.4 GHz) and 4 GB memory.

4.1 The impact of number of queries

This section looks into how the number of queries may affect the performance of the system. In the experiment, we set $|O|$ = 50,000 and varied $|Q|$ from 1,000 to 12,000. The performance results are shown in Figure 9, 10 and 11. We observe that CRQIBV outperforms CES in terms of processing time (combined query evaluation time and index update time) and storage cost. We also observe that the performance of both indexes degrades as $|Q|$ increases. However, CRQIBV is degrading at a slower rate. The reason behind this can be explained as follows. Given an object o, it is likely to be covered by more queries as $|Q|$ increases. Thus, the movement of o incurs more query result list maintenance costs (i.e., insert (or remove) object id into a query result list). We also note that the storage cost of CRQIBV is much lower than that of CES. This means the usage of bit vector can indeed achieve storage effectiveness.

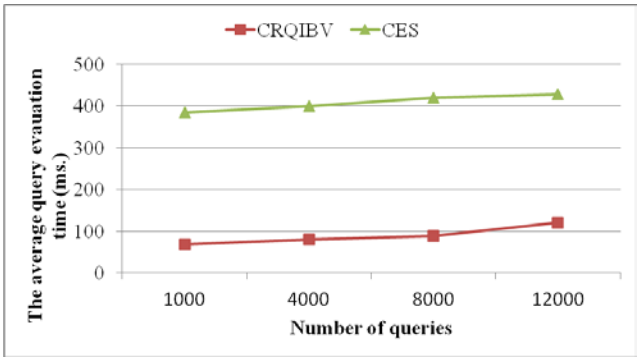


Figure 9 Query evaluation time v.s. #Queries.

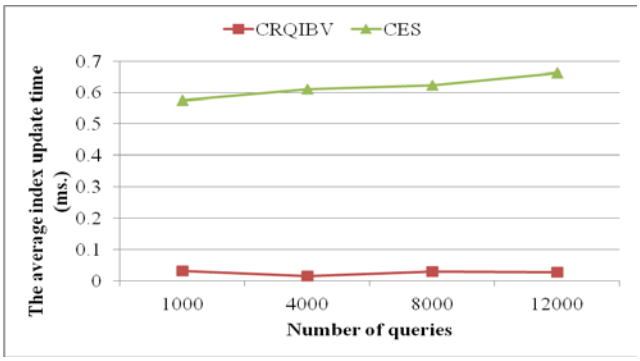


Figure 10 Index update time v.s. #Queries.

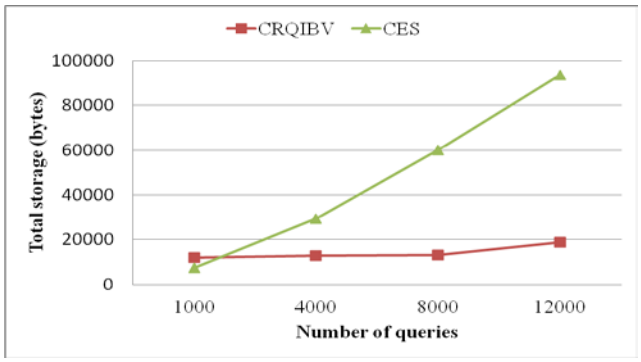


Figure 11 Memory capacity v.s. #Queries.

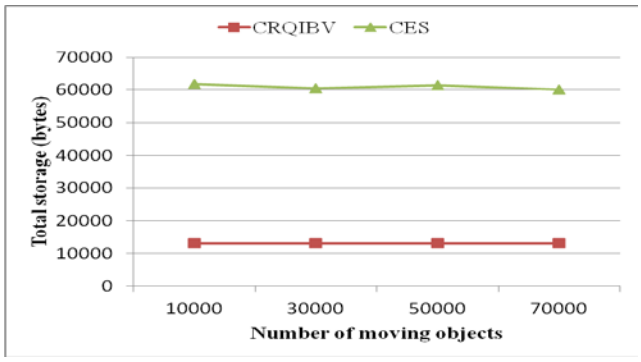


Figure 14 Memory capacity v.s. #Objects.

4.2 The impact of number of objects

Figure 12, 13 and 14 show the impact on performance as a function of object quantity. In the experiment, we fix $|Q|$ to 8,000 and varied $|O|$ from 10,000 to 70,000. Again, our CRQIBV outperforms CES under all performance metrics. In both indexes, each object location is used to search the query indexes to find all queries that cover the object. Then, the query results of the covering queries should be modified to obtain up-to-date results. As $|O|$ increases, a query q may cover more moving objects. Thus, the probability that the result of q will be modified also increases, which incurs more processing costs.

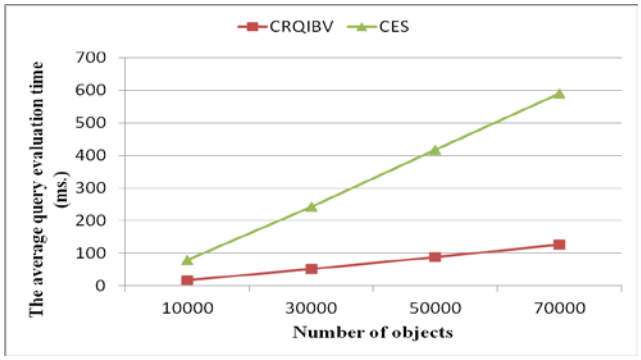


Figure 12 Query evaluation time v.s. #Objects.

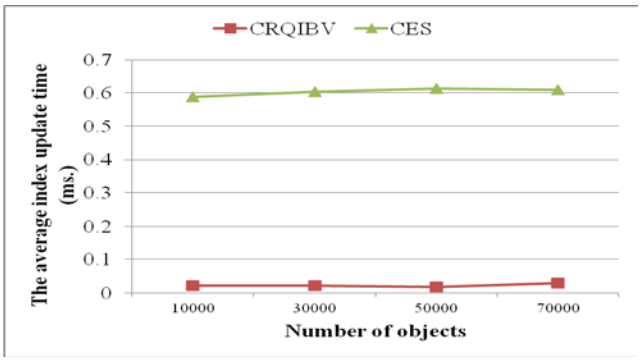


Figure 13 Index update time v.s. #Objects.

4.3 The impact of number of *QW*

In this experiment, we fix $|Q|$ = 8,000 and $|O|$ = 50,000. Four different *QW*s: 40 x 40, 80 x 80, 120 x 120 and 160 x 160 were used to evaluation the performance of two indexes. The results are shown in Figure 15, 16 and 17. As expected, CRQIBV demonstrates a superior performance compared with CES. The performance of both indexes degrades as the query size increases due to the same reason as explained in Section 4.2. We also observe that the storage cost of CES increases sharply with the increasing query size, while CRQIBV scales more gracefully. This again demonstrates that CRQIBV can be more efficient in memory usage by using bit vectors.

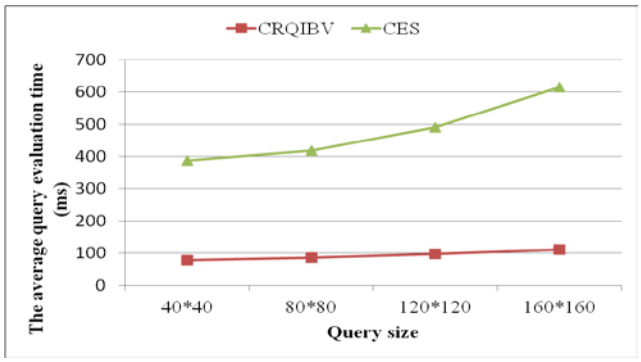


Figure 15 Query evaluation time v.s. query size.

Figure 16 Index update time v.s. query size.

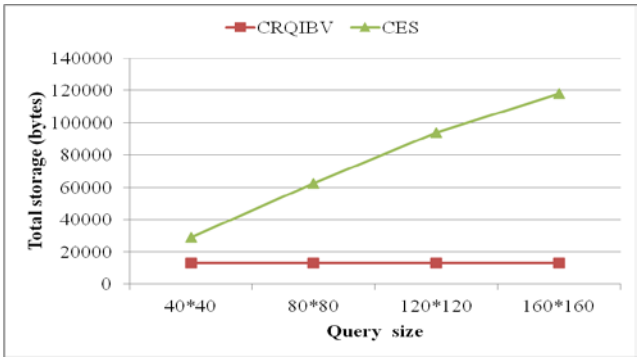


Figure 17 Memory capacity v.s. query size.

5 Conclusions

In this paper, we presented CRQIBV query index for continuous range queries processing. CRQIBV uses bit vector as the underlying data structure in order to achieve greater efficiency in some aspects. We do an initial performance comparison on CRQIBV and CES. The results show CRQIBV achieves significant performance improvement in terms of computation and storage costs. Currently, in CRQIBV, each bit vector has a fixed capacity. This may incur memory overflow if there is a burst of incoming queries. We plan to extend CRQIBV so that is can dynamically extend the capacity of a bit vector.

6 References

[1] Kun-Lung Wu, Shyh-Kwei Chen, and Philip S. Yu, "Incremental processing of continual range queries over moving objects," IEEE Transactions on Knowledge and Engineering, Vol. 18(11), Nov. 2006.

[2] D. V. Kalashnikov, S. Prabhakar, W. G. Aref, and S. E. Hambrusch, "Efficient evaluation of continuous range queries on moving objects," in Proceedings of 13th Intl. Conference on Database and Expert Systems Applications (DEXA), 2002.

[3] K. Wu, E. J. Otoo, and A. Shoshani, "Optimizing bitmap indices with efficient compression," ACM Transactions on Database Systems, Vol. 31(1), 2006.

[4] K.-L. Wu, S.-K. Chen, and P.S. Yu, "Processing continual range queries over moving objects using VCR-Based query indexes," in Proceedings of IEEE international conference of mobile and ubiquitous systems: networking and services, 2004.

[5] K.-L. Wu, S.-K. Chen, and P.S. Yu, "Efficient processing of continual range queries for location-aware mobile services," Information Systems Frontiers, Vol. 7, nos. 4-5, pp. 435-448, Dec. 2005.

[6] K.-L. Wu, S.-K. Chen, and P.S. Yu, "Shingle-based query indexing for location-based mobile e-commerce," in Proceedings of IEEE international conference of e-commerce, July 2004.

[7] S. Saltenis, C. S. Jensen, S. T. Leutenegger, and M. A. Lopes, "Indexing the positions of continuously moving objects," in Proceedings of ACM SIGMOD, pp. 331-342, 2000.

[8] L. Arge, V. Samoladas, and J. S. Vitter., "On two-dimensional indexability and optimal range searching indexing," in Proceedings of ACM PODS, pp. 346-357, 1999.

[9] C. Faloutsos and S. Roseman, "Fractals for secondary key retrieval," in Proceedings of ACM SIGMOD, pp. 47-57, 1984.

[10] N. Beckmann, H. Kriegel, R. Schneider, and B. Seeger, "The R*-tree: An efficient and robust access method for points and rectangles," in Proceedings of ACM SIGMOD, pp. 322-331, 1190.

[11] M. F. Mokbel, T. M. Ghanem, and W. G. Aref, "Spatio-temporal access methods," IEEE data engineering bullet, Vol. 26(2), pp. 40-49, 2003.

Utilizing Noise Addition for Data Privacy, an Overview

Kato Mivule
Computer Science Department
Bowie State University
14000 Jericho Park Road Bowie, MD 20715
Mivulek0220@students.bowiestate.edu

Abstract – *The internet is increasingly becoming a standard for both the production and consumption of data while at the same time cyber-crime involving the theft of private data is growing. Therefore in efforts to securely transact in data, privacy and security concerns must be taken into account to ensure that the confidentiality of individuals and entities involved is not compromised, and that the data published is compliant to privacy laws. In this paper, we take a look at noise addition as one of the data privacy providing techniques. Our endeavor in this overview is to give a foundational perspective on noise addition data privacy techniques, provide statistical consideration for noise addition techniques and look at the current state of the art in the field, while outlining future areas of research.*

Keywords: Data Privacy, Security, Noise Addition, Data Perturbation

1. Introduction

Large data collection organizations such as the Census Bureau often release statistics to the public in the form of statistical databases, often transformed to some extent, omitting sensitive information such as personal identifying information (PII). Researchers have shown that with such publicly released statistical databases in conjunction with supplemental data, adversaries are able to launch inference attacks and reconstruct identities of individuals or an entity's sensitive information [1]. Therefore while data de-identification is essential, it should be taken as an initial step in the process of privacy preserving data publishing but other methods such as noise addition should strongly be considered after PII has been removed from data sets to ensure greater levels of confidentiality [1] [2]. A generalized data privacy procedure would involve both data de-identification and perturbation as shown in *Figure 1*.

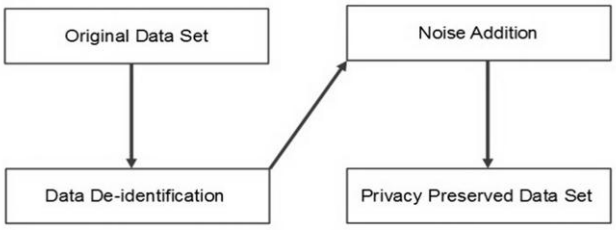


Figure 1: *Generalized Data Privacy with Noise Addition*

2. Background

In this section we take a look at some of the terms used in the noise addition procedure. *Data Privacy and Confidentiality* is the protection of an entity or an individual against illegitimate information revelation. [1]. *Data Security* is concerned with legitimate accessibility of data [2]. *Data de-identification* is the removal of personally identifiable information (PII) from a data set [3] [4]. *Data de-identification process* also referred to as *data anonymization, data sanitization,* and *statistical disclosure control* (SDC), is a process in which PII attributes are excluded or denatured to such an extent that when the data is made public, a person's identity, or an entity's sensitive data, cannot be reconstructed [5] [6]. *Statistical disclosure control* methods are classified as *non-perturbative* and *perturbative*, with the former being a procedure in which original data is not denatured, while with the latter, original data is denatured before publication to provide confidentiality [1]. Therefore de-identification of data ensures to some extent that sensitive and personal data does not suffer from *inference* and *reconstruction attacks*, which are methods of attack in which isolated pieces of data are used to infer a supposition about a person or an entity [7].

Data utility verses privacy is how useful a published dataset is to the consumer of that publicized dataset. In most instances, when publishers of large data sets do so, they ensure that PII is removed and data is distorted by noise addition techniques. However, in doing so, the original data suffers loss of some of its statistical properties even while confidentiality is granted, thus making the dataset almost meaningless to the user of the published dataset. Therefore a balance between privacy and utility needs is always sought [24] [25] [26]. Data privacy scholars have noted that achieving optimal data privacy while not shrinking data utility is an ongoing NP-hard task [27]. *Statistical databases* are non-changing data sets often published in aggregated format [28]. While data de-identification will ensure the removal of PII attributes, it has been deemed a novice method by researchers; the remaining *sanitized* data set could still be compromised and used to reconstruct an individual's identity or an entity's sensitive data [1] [2]. Therefore the remaining confidential attributes that contain sensitive information for example salary, student's GPA, need to be transformed to such an extent that they cannot be linked

with outside information in an inference attack. It is in this context that we focus on noise addition as a perturbation methodology that seeks to transform numerical attributes to grant confidentiality.

3. Related work

With an increasing interest in data privacy and security research, a number of surveys have been done articulating the progress and state of the art in the data privacy and security research field. In their survey on data privacy and security, Santos et al., present an overview on state of the art in data security techniques, placing emphasis on data security solutions for data warehousing [40]. Furthermore, in their overview, Matthews and Harel, offer a more comprehensive summary of current statistical disclosure limitation techniques, noting that that the balance between privacy and utility is still being sought with data privacy enhancing techniques [41]. Additionally Joshi and Kuo, offer an outline of state of the art data privacy techniques in Online Social Networks, in which they note how a balance is always pursued between privacy requirements for users and using private data for advertisements [42]. Yet still, in their review, Ying-hua et al., take a closer look at the current data privacy preserving techniques in data mining, providing advantages and disadvantages of various data privacy procedures [43]. While a number of current overviews on data privacy focus on the general data privacy enhancing techniques, in this paper, we focus on noise addition methods while providing statistical considerations for data perturbation.

4. Noise Addition

In this section, we take a look at noise addition perturbation methods that transform confidential attributes by adding noise to provide confidentiality. Noise addition works by adding or multiplying a stochastic or randomized number to confidential quantitative attributes. The stochastic value is chosen from a normal distribution with zero mean and a diminutive standard deviation [10] [11].

4.1. Additive Noise

Work on additive noise was first publicized by Kim [12] with the general expression that

$$Z = X + \varepsilon \quad (1)$$

Where Z is the transformed data point, X is the original data point and ε is the random variable (noise) with a distribution $e \sim N(0, \sigma^2)$. This is then added to X. The X is then replaced with the Z for the data set to be published.[13] *With stochastic noise,* random data is added to confidential attributes to conceal the distinguishing values, an example includes increasing a student's GPA by a diminutive percentage, say from 3.45 to 3.65 GPA [14]. In their work on additive noise, Domingo-Ferrer et al., outline that in additive noise, also referred to as *white noise,* concealment by additive noise anticipates that the variable of measurements x_j of the original data set X_j is continuously replaced by the variable,

$$z_j = x_j + j \quad (2)$$

Where j is the variable of normally distributed noise acquired from a random variable: $\varepsilon_j \sim N(0, \sigma_j^2)$, such that
$Cov(\varepsilon_t, \varepsilon_l)$, for all $t! = l$ thus the method preserves the mean and covariance. [20] Therefore additive noise can be expressed in a simple format as follows [21]:

$$Z = X + \varepsilon \quad (3)$$

Z is masked data value to be published, after the transformation $X + \varepsilon$. X is the original unmasked data value in the raw data set. ε (epsilon) is the random variable (noise) added to X, whose distribution is $\varepsilon \sim N(0, \sigma^2)$. Ciriani et al., note that additive noise also known as *uncorrelated noise,* preserves the mean and covariance of the original data but the correlation coefficients and variances are not sustained. Another variation of additive noise is *correlated additive noise* that keeps the mean and allows the sustenance of correlation coefficients in the original data [22].

4.2. Multiplicative Noise

Multiplicative noise is another type of stochastic noise outlined by Kim and Winkler [23] in which they describe that multiplicative noise is rendered by generating random numbers with a mean = *1*, which then are used as noise and multiplied to the original data set. Each data element is multiplied by a random number with a short Gaussian distribution, with mean = *1* and a small variance:

$$Y_j = X_j \varepsilon_j \quad (4)$$

Where Y is the perturbed data; X is the original data; E is the generated random variable (noise) with a normal distribution with mean μ and variance σ [23].

4.3 Logarithmic multiplicative noise

Kim and Winkler [23] describe another variation of multiplicative noise, in which a logarithmic alteration is taken on the original data:

$$Y_j = lnX_j \quad (5)$$

The random number (noise) is then generated and then added to the altered data [23]:

$$Z_j = Y_j + \varepsilon_j \quad (6)$$

Where X is the original data; Y is the logarithmic altered data; Z is the logarithmic altered data with noise added to it; e^x is the exponential function used to calculate the antilog.

4.4. Differential Privacy

In this section, we take a look at Differential privacy, a current state of the art data perturbation method that utilizes Laplace noise addition techniques and was proposed by Dwork (2006). Differential privacy is the latest state-of-the-art methodology in data privacy that enforces confidentiality by returning perturbed aggregated query results from databases, such that users of the databases cannot discern if particular data item has

been altered or not. This means that with the perturbed results of the query, an attacker cannot derive information about any data item in the database [33]. The database in this case is a collection of rows that represent each individual entity we seek to provide concealment. [34] According to Dwork (2008), two databases D_1 and D_2 are considered identical or similar, if they differ or disagree in only one element or row that is $D_1 \Delta D_2 = 1$. Therefore, a procedure q_n that grants confidentiality, satisfies *ε-differential privacy* if the result to any same query run on database D_1 and again run on database D_2 should probabilistically be similar, and as long as those results satisfy the following requirement: [36]

$$\frac{P[q_n(D_1) \in R]}{P[q_n(D_2) \in R]} \leq e^{\varepsilon} \qquad (7)$$

Where D_1 and D_2 are the two databases

- P is the probability of the perturbed query results D_1 and D_2 respectively.
- $q_n()$ is the privacy granting procedure (perturbation).
- $q_n(D_1)$ is the privacy granting procedure on query results from database D_1.
- $q_n(D_2)$ is the privacy granting procedure on query results from database D_2.
- R is the perturbed query results from the databases D_1 and D_2 respectively.
- e^{ε} is the exponential ε epsilon value.

Therefore to satisfy differential privacy, the probability of the perturbed query results D_1 divided by the probability of the perturbed query results D_2 should be less or equal to an exponential ε epsilon value. That is to say, if we run the same query on database D_1, and then run the same query again on database D_2, our query results should probabilistically be similar. If the condition can be mitigated in the presence or absence of the most influential observation for a particular query, then this condition will also be mitigated for any other observation. The consequence of the most dominant observation for a given query is given by Δf and assessed in the following way:

$$\Delta f = Max|f(D_1) - f(D_2)| \qquad (8)$$

For all possible realizations of D_1 and D_2, Where $f(D_1)$ and $f(D_2)$ represent the true responses to the query from D_1 and D_2 [33] [34] [35] [36]. According to Dwork (2006), the results to a query are presented as noise in the following way:

$$f(x) + Laplace(0, b) \qquad (9)$$

Where b is defined as follows for Laplace noise:

$$b = \frac{\Delta f}{\varepsilon} \qquad (10)$$

X represents a particular realization of the database, while $f(x)$ represents the true response to the query, the response would satisfy *ε-differential privacy.* The Δf must look at all possible realizations of D_1 and D_2 [33] [34] [35] [36] [37]. We could take an example in which we query the GPA of students at Bowie State University. If our Min GPA in the database is 2.0, for smallest possible GPA, and our Max GPA is 4.0 for largest possible GPA, we then calculate Δf *as* 2.0. We choose a small $\boldsymbol{\varepsilon}$ value of 0.01. The parameter b of the Laplace noise is set to $\Delta f/\varepsilon$ = 2.0/0.01 = 200. Thus we have Laplace (0, 200) noise distribution. Therefore the unperturbed results of the query + Noise from Laplace (0, 200) = Perturbed query results satisfying *ε-differential* privacy. [34] It has been noted by researchers that a smaller $\boldsymbol{\varepsilon}$ epsilon value creates greater privacy by the procedure. However, utility risks degeneration with a much smaller $\boldsymbol{\varepsilon}$ epsilon value [38]. For example, $\boldsymbol{\varepsilon}$ at 0.0001, will give $\boldsymbol{b}$ as 20000, Laplace (0, 20000) noise distribution.

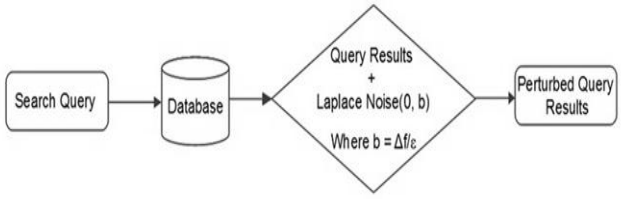


Figure 2: *A general Differential Privacy satisfying procedure.*

General steps for differential privacy shown in *Figure 2:*

- Run query on database
- Calculate the most influential observation
- Calculate the Laplace noise distribution
- Add Laplace noise distribution to the query results
- Publish perturbed query results.

4.5. Differential privacy pros and cons

Differential privacy grants across-the-board privacy, and easy to implement with SQL for aggregated data publication [39]. However, utility is a challenge as statistical properties change with a much smaller ε, as Laplace noise addition takes into account the outliers and most influential observation. [38] More noise to the data at the level of the most influential observation only renders the data useless thus balance between privacy and utility still a challenge [34] [37].

4.6. Statistical background for Noise addition

In this section, we take a look at statistical considerations for data perturbation utilizing noise addition. With noise addition, transformed data has to keep the same statistical properties as the original data. Therefore consideration has to be made for statistical characteristics such as normal distribution, mean, variance, standard deviation, covariance, and correlation

for both original and perturbed data sets.

The Mean μ, is the average of values after their total sum has been taken. In this case we would look at the summation of values then we divide them by the *n*, the quantity of values; the mathematical statement then for *the Mean* μ, is straight forward: [16]

$$\mu = \frac{1}{n} \sum_{k=0}^{n} x_i \quad (11)$$

The Normal Distribution, also known as the Gaussian distribution, used in calculating the noise addition, is a bell shaped continuous probability distribution used as an estimation to depict real-valued stochastic variables that agglomerate around a single mean. The formula for normal distribution is as follows:[15]

$$f(x) = \frac{1}{\sqrt{(2\,\pi\,\sigma^2)}} \times e^{-((x-\mu)^2/2\sigma^2)} \quad (12)$$

The parameter μ represents the mean, the point of the peak in the bell curve, while the parameter σ^2 representing the variance, the width of the distribution. The annotation $N(\mu, \sigma^2)$ represents a normal distribution with mean μ and variance σ^2. Therefore $X \sim N(\mu, \sigma^2)$ is representative of *X* distributed $N(\mu, \sigma^2)$. The distribution with $\mu = 0$ and $\sigma^2 = 1$ is cited to as the standard normal.

The Variance σ^2, in noise addition, is a measure of how data distributes itself in approximation to the mean value. The expression for variance is given by: [17]

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N} \quad (12)$$

Where σ^2 is the variance, μ is the mean, *X* being the single data values, *N* as the number of values, and $\sum (X-\mu)^2$ as the summing up of all data values *X* minus the mean μ squared.

The Standard Deviation, σ, is a measure of how distributed data is from the normal, thus we would say standard deviation is how data points are deviated from the mean. The mathematical expression is simply the square root of the variance σ^2: [18]

$$\sigma = \sqrt{\sigma^2} \quad (13)$$

Covariance: With noise addition, the measurement of how affiliated original data and perturbed data are, is crucial. Covariance, *Cov(X, Y)*, is a calculation of how affiliated the deviations between the data points *X* and *Y* are. If the covariance is positive, then the *X* and *Y* data points' inclination is to increase together, else if the covariance is negative, then the tendency is that for the two data points *X* and *Y*, one lessens while the other gains. However, if the covariance is zero, then this would signal that the data points are each autonomous. The expression for covariance is given as follows: [19]

$$Cov\,xy = \frac{1}{N} \sum x_i y_i - \overline{xy} \quad (14)$$

Correlation r_{xy} also known as the Pearson product, calculates the capability and inclination of an additive or linear relation between two data points. The correlation r_{xy} is dimensionless, autonomous of the parts in which the data points *x* and *y* are calculated [19]. If r_{xy} is = *-1*, then r_{xy} indicates a negative linear relation between the data points *x* and *y*. If $r_{xy} = 0$, then the linear relation between the two data points *x* and *y* does not exist, however, a regular nonlinear relation might exist. If $r_{xy} = +1$, then there is a strong linear relation between *x* and *y*. The expression used for correlation is: [19]

$$Correlation = r_{xy} = \frac{Cov\,xy}{\sigma_x \sigma_y} \quad (15)$$

4.7. Signal Noise Ratio (SNR)

In this section, we take a look at SNR in relation to data perturbation using noise addition, with the aim that SNR could be employed to achieve optimal data utility while preserving privacy, by measuring how much noise we need to optimally obfuscate data. In electronic signals, SNR is used to calculate a signal tainted by noise by approximating the signal power to noise power ratio, basically the ratio of the power of the signal without noise over the power of the noise.

$$SNR = \frac{Signal\,Varince}{Noise\,Variance} \quad (16)$$

With data perturbation, we could further borrow from the definition of SNR employed in Image Processing were the ratio of mean to standard deviation of a signal is used, and typically SNR is computed as the ratio of the mean pixel value to the standard deviation of the pixel values in a certain vicinity [29] [30].

$$SNR = \frac{\mu}{\sigma} \quad (17)$$

The parameter μ in this case represents the mean of the signal and the parameter σ as the standard deviation of the noise. A presumed threshold for *SNR* in image processing is based on the *Rose Criterion* which stipulates that an *SNR* of 5 is desirable in order to differentiate image details with 100 per cent confidence. Therefore, an *SNR* of less than 5 per cent will result in less than 100 per cent confidence in recognizing particulars of an image [31].

5. Illustration

In this section, we provide an example of data perturbation with noise addition for illustrative purposes. We follow a simple algorithm in implementing noise addition perturbation methodology to provide confidentiality in a published data set. The first step is the

de-identification of the data set by the removal of PII, after which we apply noise addition. In our implementation, we created a data set of 10 records for illustrative purposes and then applied the algorithm below. The original data set contained PII, we de-identified the original data set, after which we applied additive noise to the numerical attributes, and we then plotted the results in a graph, comparing the statistical properties of the original and perturbed data.

Steps for De-identification and Noise Addition

1. For all values of the data set to be published,
2. Do data de-identification
   - 2.1. Find PII
   - 2.2 Remove PII
3. For remaining data void of PII to be published,
   - 3.1. Find quantitative attributes in the data set
   - 3.2. Apply additive noise to the quantitative data values
4. Publish data set

5.1. Results of Illustration

First Name	Last Name	SSN	Age	Major	GPA	Zip code	State	Scholarship Amount	Gender
John	Artist	Xxx-xx-xxx9	23	Computer Science	4.00	21071	MA	30000.00	M
Peter	Chemist	Xxx-xx-xx10	33	Biology	3.35	31072	MD	50000.00	M
Evan	Biologist	Xxx-xx-xx11	32	Chemistry	2.19	21073	MA	25000.00	F
Joy	Music	Xxx-xx-xx12	45	History	2.99	21074	MD	67000.00	F
Eva	Pictures	Xxx-xx-xx13	23	Chemistry	3.67	21075	PA	78000.00	F
Sandra	Hollywood	Xxx-xx-xx14	21	Art	3.65	21076	MD	90888.00	F
Okello	Oscars	Xxx-xx-xx15	25	Music	4.00	21077	LA	90000.00	M
Mukisa	Grammys	Xxx-xx-xx16	30	Computer Science	2.79	21078	PA	10000.00	M
Jacinta	Historian	Xxx-xx-xx17	32	Biology	3.00	21079	CA	7000.00	F
Bosco	Activist	Xxx-xx-xx18	37	Art	2.98	21080	MO	11000.00	M

Table 1: *Original Data Set (All data for illustrative purposes).*

Age	Major	GPA	State	Scholarship Amount
23	Computer Science	3.33	MA	30000.00
33	Biology	3.35	MD	50000.00
32	Chemistry	2.19	MA	25000.00
45		2.99	MD	67000.00
23	Chemistry	3.35	PA	78000.00
21	Art	2.19	MD	90888.00
25		3.11	LA	90000.00
30	Computer Science	2.99	PA	10000.00
32	Biology	3.00		7000.00
37	Art	3.00		11000.00

Table 2*: Result after de-identification on original data.*

Scholarship Amount	Normal D of Origin Scholarship	Perturbed Scholarship Amount	Normal D of Perturbed Scholarship
30000.00	0.7602	37296.88	0.7788
50000.00	0.9859	52087.4	0.9555
25000.00	0.6312	29403.32	0.6066
67000.00	0.9997	72389.89	0.9986
78000.00	1.0000	84639.16	0.9999
90888.00	1.0000	97116.52	1.0000
90000.00	1.0000	91554.68	1.0000
10000.00	0.2172	18977.3	0.3494
7000.00	0.1575	10455.81	0.1777
11000.00	0.2398	17932.86	0.3253

Table 4: *Results of the Normal Distribution of Original Perturbed Scholarship Amount.*

		Random Noise Between 1000 and 9000	
	Scholarship Amount	Random Noise	Perturbed Scholarship Amount
	30000.00	7296.88	37296.88
	50000.00	2087.4	52087.4
	25000.00	4403.32	29403.32
	67000.00	5389.89	72389.89
	78000.00	6639.16	84639.16
	90888.00	6228.52	97116.52
	90000.00	1554.68	91554.68
	10000.00	8977.3	18977.3
	7000.00	3455.81	10455.81
	11000.00	6932.86	17932.86
Mean	***20500***		***27614.87***
Standard Deviation	***13435.0288425444***		***13692.4298530319***
Variance	***180500000***		***187482635.2802***

Table 3: *Random noise between 1000 and 9000 added to Scholarship attribute.*

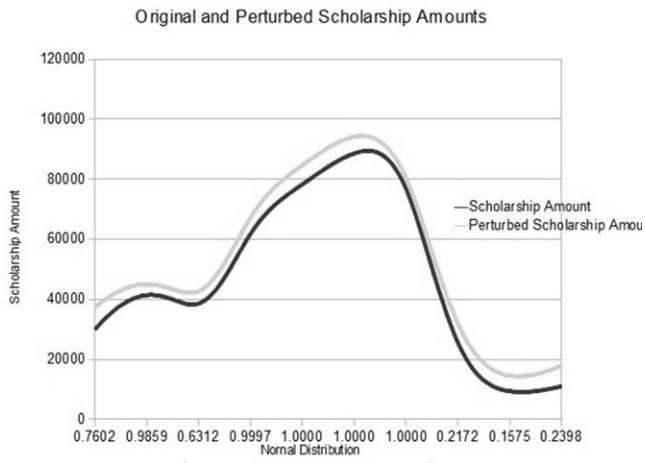


***Figure 3**: Results of the normal distribution of original and perturbed scholarship amount*

Covariance between Original Scholarship Data set and Perturbed Scholarship Data set = *1055854875.465***.** Since Covariance is positive, it shows that the two data sets move together in the same direction. *Correlation* between Original Scholarship Data set and Perturbed Scholarship Data set = *0.999***.** Since Correlation is a strong positive, it shows a relationship between the two data sets, increasing and decreasing together.

6. Conclusion

We have taken a look at data perturbation utilizing noise addition as a methodology used to provide privacy for published data sets. We also took a look the statistical considerations when utilizing noise addition. We provided an illustrative example showing that de-identification of data when done in concert with noise addition would add more to the privacy of published data sets while maintaining the statistical properties of the original data set. However, generating perturbed data sets that are statistically close to the original data sets is still a challenge as consideration has to be made for the tradeoff between utility and privacy; the more close the perturbed data is to the original, the less confidential that data set becomes, and the more distant the perturbed data set is from the original, the more secure but then, utility of the data set might be lost when the statistical characteristics of the origin data set are lost. Noise generation certainly affects the level of perturbation on the original data set. Yet still, striking the right balance between privacy and utility remains a factor. While state of the art data perturbation techniques such as differential privacy provide hope for achieving greater confidentiality, achieving optimal data privacy while not shrinking data utility is an ongoing NP-hard task. Therefore more research needs to be done on how optimal privacy could be achieved without degrading data utility. Another area of research is how noise addition techniques could be optimally applied in the cloud and mobile computing arena, given the ubiquitous computing era.

7. References

[1] V. Ciriani, et al, 2007. Secure Data Management in Decentralized System, Springer, ISBN 0387276947, 2007, pp 291-321.

[2] D.E Denning and P.J Denning, 1979. Data Security, ACM Computing Surveys, Vpl. II, No. 3, September 1, 1979.

[3] US Department of Homeland Security, 2008. Handbook for Safeguarding Sensitive Personally Identifiable Information at The Department of Homeland Security, October 2008. [Online]. Available at: http://www.dhs.gov/xlibrary/assets/privacy/privacy_guide_spii_handbook.pdf

[4] E. Mccallister and K. Scarfone, 2010. Guide to Protecting the Confidentiality of Personally Identifiable Information (PII) Recommendations of the National Institute of Standards and Technology, NIST Special Publication 800-122, 2010.

[5] S.R. Ganta, et al, 2008. Composition attacks and auxiliary information in data privacy, Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - SIGKDD '08, 2008, p. 265.

[6] A. Oganian, and J. Domingo-Ferrer, 2001. On the complexity of optimal microaggregation for statistical disclosure control, Statistical Journal of the United Nations Economic Commission for Europe, Vol. 18, No. 4. (2001), pp. 345-353.

[7] K.F. Brewster, 1996. The National Computer Security Center (NCSC) Technical Report - 005V olume 1/5 Library No. S-243,039, 1996.

[8] P. Samarati, 2001. Protecting Respondent's Privacy in Microdata Release. IEEE Transactions on Knowledge and Data Engineering 13, 6 (Nov./Dec. 2001): pp. 1010-1027.

[9] L. Sweeney, 2002. k-anonymity: A Model for Protecting Privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10, 5 (Oct. 2002): pp. 557-570.

[10] Md Zahidul Islam, Privacy Preservation in Data Mining Through Noise Addition, PhD Thesis, School of Electrical Engineering and Computer Science, University of Newcastle, Callaghan, New South Wales 2308, Australia, November 2007

[11] Mohammad Ali Kadampur, Somayajulu D.V.L.N., A Noise Addition Scheme in Decision Tree for, Privacy Preserving Data Mining, JOURNAL OF COMPUTING, VOLUME 2, ISSUE 1, JANUARY 2010, ISSN 2151-9617

[12] Jay Kim, A Method For Limiting Disclosure in Microdata Based Random Noise and Transformation, Proceedings of the Survey Research Methods, American Statistical Association, Pages 370-374, 1986.

[13] J. Domingo-Ferrer, F. Sebé, and J. Castellà-Roca, "On the Security of Noise Addition for Privacy in Statistical Databases," in Privacy in Statistical

Databases, vol. 3050, Springer Berlin / Heidelberg, 2004, p. 519.

[14] Huang et al, Deriving Private Information from Randomized Data, Special Interest Group on Management of Data - SIGMOD 2005 June 2005.

[15] Lyman Ott and Michael Longnecker, An introduction to statistical methods and data analysis, Cengage Learning, 2010, ISBN 0495017582, 9780495017585, Pages 171-173

[16] Martin Sternstein, Barron's AP Statistics, Barron's Educational Series, 2010, ISBN 0764140892, Pages 49-51.

[17] Chris Spatz, Basic Statistics: Tales of Distributions, Cengage Learning, 2010, ISBN 0495808911, Page 68.

[18] David Ray Anderson, Dennis J. Sweeney, Thomas Arthur Williams, Statistics for Business and Economics, Cengage Learning, 2008, ISBN 0324365055, Pages 95.

[19] Michael J. Crawley, Statistics: an introduction using R, John Wiley and Sons, 2005, ISBN 0470022973, Pages 93-95.

[20] J. Domingo-Ferrer and V. Torra (Eds.), On the Security of Noise Addition for Privacy in Statistical Databases, LNCS 3050, pp. 149–161, 2004.# Springer-Verlag Berlin Heidelberg 2004.

[21] Ruth Brand, Microdata Protection Through Noise Addition, LNCS 2316, pp. 97–116, 2002. Springer-Verlag Berlin Heidelberg 2002.

[22] [22] Ciriani et al, Microdata Protection,Secure Data Management in Decentralized System, pages 291-321, Springer, 2007.

[23] Jay J. Kim and William E. Winkler, Multiplicative Noise for Masking Continuous Data, Research Report Series, Statistics #2003-01, Statistical Research Division, U.S. Bureau of the Census.

[24] Rastogi et al, The boundary between privacy and utility in data publishing, VLDB ,September 2007, pp. 531-542.

[25] Sramka et al, A Practice-oriented Framework for Measuring Privacy and Utility in Data Sanitization Systems, ACM, EDBT 2010.

[26] Sankar, S.R., Utility and Privacy of Data Sources: Can Shannon Help Conceal and Reveal Information?, presented at CoRR, 2010.

[27] Wong, R.C., et al, Minimality attack in privacy preserving data publishing, VLDB, 2007. pp.543-554.

[28] Adam, N.R. and Wortmann, J.C., A Comparative Methods Study for Statistical Databases: Adam and Wortmann, ACM Comp. Surveys, vol.21, 1989.

[29] Jeffrey J. Goldberger, Practical Signal and Image Processing in Clinical Cardiology, Springer, 2010, Page 28-42

[30] John L. Semmlow, Biosignal and biomedical image processing: MATLAB-based applications, Volume 22 of Signal processing and communications CRC Press, 2004, ISBN 9780824750688, Page 11.

[31] Jerrold T. Bushberg, The essential physics of medical imaging, Edition 2, Lippincott Williams & Wilkins, 2002, ISBN 0683301187, 9780683301182, Page 278-280.

[32] Narayanan, A. and Shmatikov, V., 2010. Myths and fallacies of "personally identifiable information". In *Proceedings of Commun. ACM. 2010*, 24-26.

[33] Dwork, C., Differential Privacy, in ICALP, Springer, 2006

[34] Muralidhar, K., and Sarathy, R., Does Differential Privacy Protect Terry Gross' Privacy?, In Privacy in Statistical Databases, Vol. 6344 (2011), pp. 200-209.

[35] Muralidhar, K., and Sarathy, R., Some Additional Insights on Applying Differential Privacy for Numeric Data, In Privacy in Statistical Databases, Vol. 6344 (2011), pp. 210-219.

[36] Dwork, C., Differential Privacy: A Survey of Results, In Theory and Applications of Models of Computation TAMC , pp. 1-19, 2008

[37] M. S. Alvim, M. E. Andrés, K. Chatzikokolakis, P. Degano, and C. Palamidessi, "Differential privacy: on the trade-off between utility and information leakage," Aug. 2011. [Online]. Available: http://arxiv.org/abs/1103.5188

[38] Fienberg, S.E., et al, Differential Privacy and the Risk-Utility Tradeoff for Multi-dimensional Contingency Tables In Privacy in Statistical Databases, Vol. 6344 (2011), pp. 187-199.

[39] A. Haeberlem, B.C. Pierce, and A. Narayan, "Differential privacy under fire," in Proceedings of the 20th USENIX Security Symposium, Aug. 2011.

[40] Santos, R.J.; Bernardino, J.; Vieira, M.; , "A survey on data security in data warehousing: Issues, challenges and opportunities," EUROCON - International Conference on Computer as a Tool (EUROCON), 2011 IEEE , vol., no., pp.1-4, 27-29 April 2011

[41] Joshi, P.; Kuo, C.-C.J.; , "Security and privacy in online social networks: A survey," Multimedia and Expo (ICME), 2011 IEEE International Conference on , vol., no., pp.1-6, 11-15 July 2011

[42] Matthews, Gregory J., Harel, Ofer, Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy, Statistics Surveys, 5, (2011), 1-29 (electronic).

[43] Liu Ying-hua; Yang Bing-ru; Cao Dan-yang; Ma Nan; , "State-of-the-art in distributed privacy preserving data mining," Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on , vol., no., pp.545-549, 27-29 May 2011

A Framework for Re-optimizing Repetitive Queries

Feng Yu, Wen-Chi Hou
Department of Computer Science
Southern Illinois University
Carbondale, IL 62901
{fyu, hou}@cs.siu.edu

Cheng Luo
Department of Mathematics and Computer Science
Coppin State University
Baltimore, MD 21216-3698
cluo@coppin.edu

ABSTRACT

In this paper, we develop a comprehensive framework forre-optimization of a large and useful set of queries, called repetitive queries. Repetitive queries refer tothosequeries that are likely to be used repeatedly or frequently in the future. Theydeserve more optimization efforts than ordinary ad hoc queries. In this research, weidentify statistics, called sufficient statistics,that are sufficient to compute the exact frequency distributions of the intermediate results of all plans ofa query. We present two innovative techniques to conductre-optimization, an eager and a lazy re-optimization. The eager approach gathersall the sufficient statistics for a query at once and generates the best plan. The lazy re-optimization gathers only the statistics that are needed to correctlarge estimation errors found in the plan and generates a revised plan. We further adapt the two basictechniques to constantly changing environmentsby continuously monitoring and revising the plans, called adaptive re-optimization. The adaptive re-optimization is devised to detectand remedy potential sub-optimality in the plans in a timely manner for the entire lifetime of the query. Our work realizes the promise made by the query optimizers, namely, executing queries in the optimal fashions, at least for the repetitive queries.

1. INTRODUCTION

A query optimizer generally uses statistics on databases[7,9, etc.] and assumptions aboutattribute values [3, 11] to estimate the cost of alternative plans and selects the best foris important for an optimizer ofto find the most efficient plans because studies [3, 6] have shown that executions with sub-optimal plans can be orders of magnitude slower than with the optimal ones. Unfortunately, due to the sufficiency of statistics stored in the database and the validityassumptions made, query optimizers often cannot findtheplan that is truly the bestin their search spaces for the queries. Thus, some database systems, like Sybase and Oracle, allow users to force the join orders; some, e.g., Sybase, even allow users to explicitly edit the plans [10]. Unfortunately, such measures cannot guarantee success and canbecumbersome and slow for complex queries.

Query re-optimization aims to refinethe execution plans of queries. It can havea tremendous impact on the performance ofsystems. Unfortunately,there hasn't been much work on this subject yet. In the literature, some[1, 5, 6, 8] have focused on refiningexecution plans of ongoing queries on-the-fly, while others [12, 2] focused on refiningcost estimation offuture queries using statistics collected in the previous executions. In this paper, we are interested in refiningcost estimation and execution plans for future queries, similar to the latter.

Stillger et al. [12] collected cardinality information from queries and used it to adjustthe cost estimation offuture queries. Unfortunately, statistics, like cardinalities and selectivities, obtained from one plan may not besufficient for estimation of another planbecause as the orders ofjoins change, the inputs ofthe joins and the outputs thereof are allchanged; certainly, they can hardly be sufficient for estimation of different queries either. The situation is further exacerbated by adding other operators, such as selects, projects, unions, and differences, to the queries. It can require prohibitivelylarge amounts of statistics, if ever possible,to compute the intermediate results' sizes for all possible plans of allqueries. Therefore, in this research, we set a more realistic goal by restricting ourselves to optimizing a subset, but a large and useful subset, of queries, called repetitive queries.

There are also many useful queries, such as those usedfor generating periodic reports, performing routine maintenances, summarizing and grouping data for analysis, etc., that are run frequently, periodically, or repeatedly, hereby called repetitive queries. They areoften stored in the database for convenient re-uses for the long term. They can constitute a large portion of the daily activities. Any sub-optimality in the execution plans of repetitive queries could mean repetitive and continued wastes ofsystem resources and time in the future. Repetitive querieshave profound impacts on the performance of systems and deserve more optimization efforts than ad hoc queries.

In this research, weidentify statistics, called sufficient statistics,that are sufficient to compute the sizes ofintermediate results of all plans of a query.We consider queries withselect, bag/set project, Cartesian product, bag/set union, and bag/set difference operations

in the forms of linearas well as bushy trees [4, 10]. The relations can be bags as well as sets of tuples. The sufficient statistics can be gathered either on-line at runtime or off-line in spare time. The gathered statistics can be used by the(existing) query optimizer to findthe best plan in its search space, conveniently called the optimal plan here.We present two innovative methods to conductquery re-optimization,an eager and a lazy approach. The eager approach gathers allsufficient statistics at once andderives the optimal plans. The lazy re-optimization gathers onlythe statistics that are needed to correctlarge cardinality estimation errors found in the plan and generates a revised plan. We adapt the two basic re-optimization techniques to constantly changing database environmentsby continuously monitoring the executions and revising the plans,called adaptive re-optimization.

Our work distinguishes itself fromothers in the area of query re-optimization. We present a comprehensivesolution tothe re-optimization of repetitive queries. We cover a large class of queries that includes the select, project, Cartesian product, union and difference operations on bags and sets of tuples in the forms of linear and bushy trees. The identification of the sufficient statistics makes it entirely possible to compute the exact intermediate results of all plans of a query and hence the re-optimization.We propose several innovative ways to conduct re-optimizations, realizing the ultimate goal of query optimization for repetitive queries.

The rest of the paper is organized as follows. In Section 2, we briefly review papers related tothe query re-optimization. In Section 3, we propose a re-optimization framework and give a functional overview of each component of it. In Section 4, we identify statistics that are sufficient to compute the intermediate results of all plans of a query. In Section 5, we present different ways to collect the sufficient statistics. In Section 6, we proposeseveral innovative ways to conduct re-optimization. Section 7 is the conclusions and future work.

2. LITERATURE SURVEY

The work inquery re-optimization can be classified into two categories: (1) re-optimizations ofongoing queries and(2) re-optimization of future queries. Our work falls into the second category.

Most re-optimization work [1, 5, 6] belongs to the first category. Kabra et al. [5]collected statistics during the execution of a query and used them to optimize the rest of the execution by either changing the execution plan or improving the resource allocations. Heuristics were used to determine if the benefits of re-optimization outweighed the overheads. Markl et al [6] computed the validity range of each plan. When the actual result falls outside the validity range, a re-optimization is triggered. Intermediate results are saved for potential re-uses in the new join order. Instead of computing a point estimate ofthe cardinality of an operator, Babu, et al [2] used interval estimates to account for the uncertainty of the estimation. Within the bounding box, they selected robust and switchable plans that avoid re-optimization and loss of pipelined work done earlier.

[2, 12] fall in the second category. Stillger et al. [12] collectedcardinalities ofoperatorsand used them to adjustselectivity estimationsof futurequeries. It may work well if the queries beara strong resemblance to the previous ones. Chaudhuri, et al. [2] useddistinct page count, as opposed tothe cardinality, as the cost measure for an operator. The page count accounts for the clustering effect of data on the disks, reflecting the computation cost or time in a more direct way.

3. RE-OPTIMIZATION FRAMEWORK

In Figure 3.1, we outline the re-optimization framework with solid lines and arrows highlighting the new components and paths added to the database system while the dotted lines the existing components and paths.Note that we do not intend to modifythe existing query optimizers, but justto provide them with sufficient statistics tofindthe execution plans that are trulythe best in their search spaces.

When a repetitive query is being executed, statistics can be collected, as indicated by the on-line statistics gathering box in the figure. One can also gatherstatistics off-line whenever it is convenient, especially in spare time, as indicated by the off-line statistical gathering box. The on-line approach takes advantage of the query evaluation to gather readily available statistics while the off-line approach gathers the sufficient statistics without interfering with the executions of queries. Section 5 has more details on these approaches.

The re-optimizer supplies the statistics gathered to the query optimizer to generate the optimal or a revised plan. Depending upon the amounts of statistics supplied to the optimizer, the re-optimization can be conducted either in an eager or a lazy fashion. The eager re-optimization refers to the situations where all sufficient statistics of a query are provided to the optimizer at once to generate the optimal plan. As for the lazy re-optimization, only selected statistics that are likely to correct large estimation errors found in the plan are gathered and provided to the optimizer to generate a revised plan. It may take a few cycles to arrive at the optimal plan. The optimal or revised plan is stored in the system for subsequent executions of the query.

The adaptive re-optimization is devised to maintain the optimality of plans throughout the lifetime of the queries. By continuously monitoring the executions and, if necessary, modifying the plans, optimality can be

maintained even though the underlying database has undergone substantial changes. The adaptive re-optimization can be accomplished by using either the eager or the lazy re-optimization method. Section 6 has detailed discussions on these re-optimization methods.

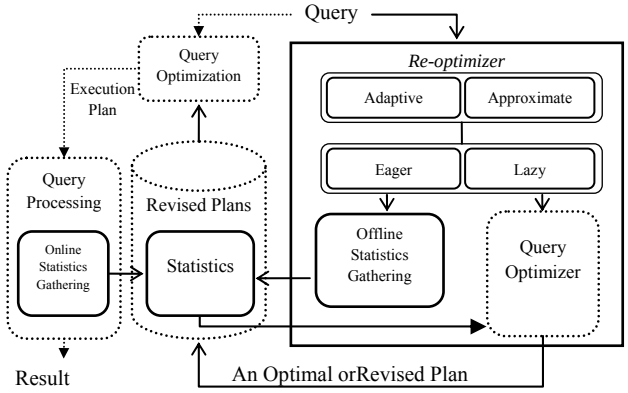


Figure 3.1 Re-optimization Architecture

Sincethe re-optimization is mainly an off-lineprocess, it can afford to use more time and resources to search for a better plan.Therefore, it may be worth employing a more sophisticated query optimizer that searches a larger solution space. Note that the proposed sufficient statistics are sufficient for computingthe exact intermediate results' sizes of all plans of a query derived by using the commonly used algebraic laws and optimization heuristics.

4. SUFFICIENT STATISTICS

We assume relational algebraic laws, such as the commutative and associative laws for joins, Cartesian product, and unions, and useful heuristics, such as pushing of the selections and projections down the tree, are used to generate alternative plans. Due to space limitation, readers are referred to [13] for proofs of all lemmas and theorem.

4.1 Common Properties of Execution Trees

First, we identify some useful properties of the plans generated. Consider a university database with three relations: Assignment(course_id,tname, dept), Books (book_id, title, publisher), and CoursesText(course_id, book_id), with their keys underlined. We assume a course can use more than one textbook. Consider the query: Print the titles of the books published by "PH" and used in the courses taught by teachers in the CS department. Figure 3.1 show two alternative plans of the query in which selections and projections have been pushed down.

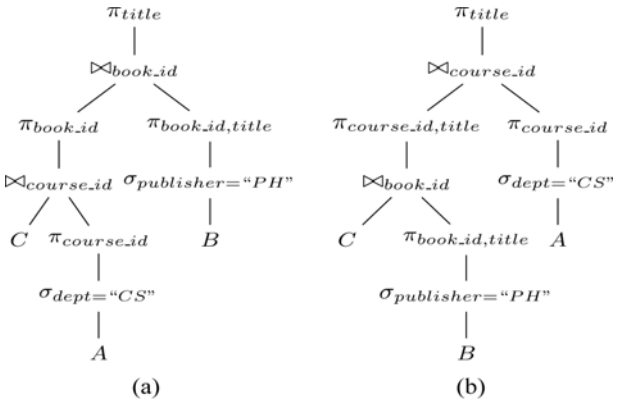


Figure 3.1 Execution Plans

In the figure, we have used, for simplicity, A for Assignment, B for Books, and C for CourseText. The two plans have different join orders.

Example 1. In Figure 3.1 (a) and (b), operand relations A, B, and C each are preceded by the same select operation in both plans. That is, in both plans, A is preceded by $\sigma_{dept="CS"}$, (i.e. $\sigma_{dept="CS"}(A)$), B by $\sigma_{publisher="PH"}$ (i.e. $\sigma_{publisher="PH"}(B)$), and C by no selection . □

LEMMA 1. Each operand relation is preceded by the same selection conditions, if any, in all plans of the query.

Definition. Select-modified relation. Let R be an operand relation of a query. After selections have been pushed as far down the tree as possible, the select-modified relation of R, denoted by R', refers to R and its immediate preceding selection, if any, in the tree. If R has no selection preceding it, R itself is the select-modified relation of R. □

4.1 Sufficient Statistics

Let attr(mr) be the set of attributes of the modified relation "mr", and basis(mr) be the set of basis attributes for "mr". We assume the highest node representing the modified relation has been marked as a modified relation. For example, for the select-modified relation $\sigma_{dept="CS"}(A)$,we assmue the node $\sigma_{dept="CS"}$ has been marked as a modified relation.

AlgorithmDist_Domain_Bases (node, bases)
{
if (node is not a modified relation)
{ for (each childnode of node)
Dist-Domain_Bases (childnode, bases);
switch (type of node)
{
Case root:
for (each modified relation mr)
{ basis(mr)=bases ∩ attr(mr);

```
if (basis(mr)=∅) basis(mr)={count};
            }
              Case select:
bases = bases∪ {attributes in the
selection condition};
              Case set projection:
bases =bases∪ {attributes on the
projection list };
              Case difference:
bases =bases∪ {keyattributes in the
operandrelations};
              Case set union:
bases =bases∪ {keyattributes in the
operandrelations};
          }
      }
}
```

Figure 4.2 Attributes of Frequency Distributions

Example 2.Consider the plans in Figure 3.1(a) and (b) again and let the projections in the trees now be the set projections. For modified relation A', basis(A') = {course_id} because course_id is a join attribute and also an attribute of the set projection π_{course_id}. For B', basis{B'}={book_id, title} because book_id is both a join and a set project attribute while title appears in a set projection. Notice that title would not have been included in basis{B'} if the projections were the bag projections. For C', basis(C')={course_id, book_id} because course_id is a join attribute and book_id is both a join and a set project attribute. □

Again, in Example 2, one can use either the plan in Figure 3.1(a) or (b) to find the bases of the frequency distributions' domains and the results are the same. The following Lemma gives a formal proof of such phenomenon under the assumption that selections and projections are pushed as far down the trees are possible in all plans.

LEMMA 2. The algorithm Dist_Domain_Bases() derives the same bases for all plans of a query.

Example 3.Consider the plan in Figure 3.1(a) with all projections being the set projections. As discussed in Example 2, basis(A')={course_id}, basis(B')={course_id, title}, and basis(C')= {course_id, book_id}. Let $f_{A'}(course_{id}), f_{B'}(book\_id, title)$, and $f_{C'}(course\_id, book\_id)$ be the frequency distributions constructed for modified relations A', B', and C', respectively.

Let $A'' = \pi_{course\_id}(A')$. The frequency distribution of A'' (= $\pi_{course\_id}(A')$), denoted by $f_{A''}(course\_id)$, can be computed by first coalescing the frequency values of $f_{A'}(course\_id)$ on attribute course_id. Since both $f_{A'}$ and $f_{A''}$ are functions on course_id, the coalescing really has no effect. As for the duplication elimination function of the set projection, we just need to set any frequency values that are greater than 1 to 1 to reflect the effect.

Let $B'' = \pi_{book\_id,title}(B')$. The frequency distribution of B' (= $\pi_{book\_id,title}(B')$), denoted as $f_{B''}(book\_id, title)$, can be computed in a similar way. The coalescing of frequencies on attributes book_id and title has no effect on the frequencies because $f_{B'}$ and $f_{B''}$ are all defined on the same set of attributes {book_id, title}. To reflect the duplicate elimination of the set projection, any frequency values greater than 1 need to be set to 1. Then, we can use $f_{C'}(course\_id, book\_id)$ and $f_{A''}(course\_id)$ to derive the resulting frequency distribution of $C' \bowtie_{couse\_id} A''$, denoted by $f_{C'\bowtie A''}(course\_id, book\_id)$, as

$$f_{C'\bowtie A''}(c,b) = f_{C'}(c,b) \times f_{A''}(c). \quad (1)$$

for given c (course_id) and b (book_id) values. To compute the distribution of $\pi_{book\_id}(C' \bowtie_{\text{course\_id}} A'')$ with duplicates deleted, denoted by $f_{\pi(C'\bowtie A'')}(book\_id)$, we first compute

$$f_{\pi(C'\bowtie A'')}(b) = \sum_{c\in Dom(cours\_id)} f_{C'\bowtie A''}(c,b), \quad (2)$$

for given course_id c and book_id b values to coalesce the frequencies on book_id and then set any frequency values that are greater than 1 to 1 to reflect the effect of the duplicate elimination of the set projection. Finally, $f_{\pi(C'\bowtie A'')}(book\_id)$ can be used with $f_{B''}(book\_id, title)$ to derive the frequencies of the final join result as

$$f_{\pi(C'\bowtie A'')\bowtie B''}(b,t) = f_{\pi(C'\bowtie A'')}(b) \times f_{B''}(b,t), \quad (3)$$

for given title t and book_id b. The attribute title will be used to coalesce the frequencies and any frequency values greater than 1 are set to 1 to reflect the last set projection on the attribute title. □

To derive the resultant frequency distributions of a bag union, one simply adds the corresponding input frequencies; for a Cartesian-product, one multiplies the frequencies of each pair of attribute values, one from a different relation.

THEOREM. The proposed sufficient statistics for a query are sufficient to compute the frequency distributions of the intermediate and final results of all plans of the query.

5. STATISTICAL GATHERING

Generally speaking, one can either gather the statistics while the query is being executed (i.e., an on-line approach), or whenever it is not interfering with the execution of the query (i.e., an off-line approach).

5.1 On-line Statistical Gathering

As tuples flow through the select operator of a modified relation, we construct the frequency distribution of the modified. Figure 5.1 shows the points where the sufficient statistics are gathered.

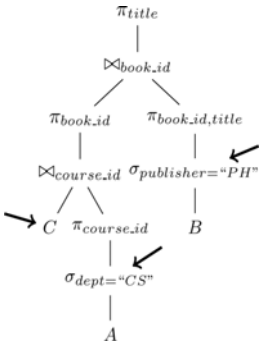


Fig. 5.1 Statistics Gathering Points

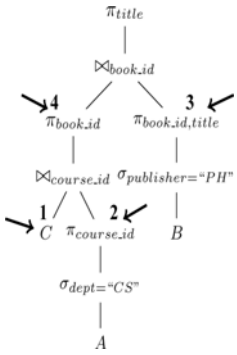


Fig. 6.1 Cardinality Checkpoints

5.2 Off-line Statistics Gathering

Instead of gathering statistics at runtime, one can collect the statistics off-line insystem's spare time or whenever the DBA feels appropriate. This approach does not interference with the executions of the queries and thus has no runtime overheads. However, re-evaluations of the modified relations would benecessary.

6. RE-OPTIMIZATION

In this section, we discuss different ways to attain the optimal plans. In general, an optimal plan can be obtained in one step or multiple steps, which are called an eager and a lazy re-optimization, respectively. We will also discuss other more sophisticated methodsthat are built on the tops of the two fundamental methods. It is noted thata re-optimization is mainly an off-line process. It can be conducted whenever appropriate, for example, at system's spare time.

6.1 Eager Re-optimization

The eager re-optimizationis probably the most straightforward way to conducta re-optimization. For a given query, wefirst collect all the sufficient statisticsand then provide them to the query optimizer to search for the optimal plan. The sufficient statistics can be obtained by any of the statistics gathering methods mentioned in Section 5 and discarded after use. The optimal plan is stored in the system for subsequent executions of the query.

6.2 Lazy Re-optimization

The plans generated using the conventional statistics stored in the database catalog maysometimes generate plans of good quality. Inthe lazy re-optimization, we attempt to use the conventional statistics in the catalog as much as possible unless large estimation errors (on the intermediate results' cardinalities (sizes)) have been found in the plan (to be discussed shortly); and only then is the statistics gathering process invoked. Note that we only gather the statistics that are likely tocorrect the large estimation errors found in the plan, not the entire set of the sufficient statistics. This lazy approach tries only to replace the statistics in the catalog that are not accurate enough for uses; itcan alsospread the cost of statistics gathering over a longer period of time.

We intend to monitorthe cardinalities (not the frequency distributions) ofintermediate results during the execution to see if any large estimation errors have occurred. In Figure 6.1, a redraw of Figure 4.1(a),the arrows indicate the checkpoints where cardinalities are to be monitored.

The gathering of cardinalities, done by counting the numbers of tuples flowing through the checkpoints, is simple and should incur no noticeable overhead. By comparing the actual cardinalities against the estimated ones, large estimation errors can be identified. Here, the estimated cardinalities refer to the cardinalities calculated by the optimizer when selecting the plans.

A threshold (e.g., 5%, to be discussed in Section 6.2.3) is set up to determine if any of the estimation errors is large enough to warrant the gathering of accurate statistics. We shall, in the next subsection, discuss how to identify modified relations for which accurate statistics shouldbe gathered to correct the large estimation errors. The desiredstatistics can be gathered by the off-line approach discussed in Section 5.

The optimizer uses the newly gatheredaccurate statistics, supplemented with the statistics in the catalog, to derive a revised plan. If there are still large estimation errorsfound in the revised plan in subsequent executions, more statistics are to be gathered. In order to reuse previously gathered statistics, the statistics need to be stored in the systemwith the query. It can be observed that in the worst case, all the sufficient statistics will eventually be collected to attain the optimal plan.

6.2.1 Placement of Checkpoints

Checkpoints can be placed essentiallyat every place in the plan. While there may be many good schemes to place checkpoints, here, we discuss a simple one that places checkpoints only at the inputs of binary operators of the plan, as shown in Figure 6.1. One reason is that the inputs of an operatordetermine the output of it. Moreover, the inputs of a binary operation often have been modified by a series ofunary operations likeselections and projections, such asthe input of the join$\bowtie_{\text{course_id}} \pi_{course_id}\left(\sigma_{dept="CS"}(A)\right)$in Figure 6.1.It is

very difficult for an optimizer to get good estimates of the inputs using the statistics stored in the catalog. Consequently, the inputs are good places to catch potential estimation errors.

6.2.2 Gathering Selected Statistics

When large estimation errors are found, we need to identify the modified relations for whichaccurate statistics canlikelycorrect the estimation errors.

The inaccuracies in the estimations of cardinalities anddata distributions of the input relations are the two main factors contributing to the errors. The checkpoints placed between the binary operators andthe modified relations, e.g., checkpoints at 1, 2, and 3 in Figure 6.1, serve just the purpose of detectingcardinality estimation errors on the inputs of the operators. Thus, if large estimation errors have occurred at such places,the accurate statistics for the respective modified relations should be gathered.

Inaccuracies in the input data distributions estimation are another factor contributing to the estimation errors. For example, dependent upon howthe join attribute values of the input tuples match, the results of a join can be quite different even though the inputs are of fixed sizes.To detect estimation errors caused by the lack of information or inappropriate assumptions on the data distributions, one has to rely on the checkpoints placed above the binary operators, such as the checkpoint 4 in Figure 6.1. Unfortunately, such checkpoints cannotpinpoint exactly the modified relation(s) for which accurate statistic can correct the estimation errors foundin the plan because inaccuracies inany of modified relations below the operators cancontribute to the errors. Therefore, we can only select modified relations for which accurate statistics can most likely correct the errors.

We summarize the above discussions to provide the followingheuristics to determine the modified relations for which accurate statistics should be collected.

Heuristics:

1) if a large estimation error has occurred between a binary operator and a modified relation (e.g., 1, 2 and 3 in Figure 6.1), gather statistics for that modified relation.

2) if a large estimation error has occurred above a binary operator (e.g., 4 in Figure 6.1) and no relation under it in the tree for which statistics has been gathered (in the current run), gather statistics for one or more modified relations at which larger errors have occurred.

In rule (2), one can be conservative to select only one modified relation, orbe aggressive to select more than one modified relation for which statistics are collected. Other heuristics are certainly possible and are left for future research.

Example 8. In Figure 6.1, if there is a large error (greater than the threshold value) found at 1 (2 or 3), then we gather statistics for the modified relation of C (A or B). If a large error is found at checkpoint 4 and no modified relation under it has been selected, then we can select either the modified relation of C or A, dependent upon where a larger error has occurred, at 1 or 2. Certainly, we canbe more aggressive to select both. □

6.2.3 Threshold

Iflarge estimation errorsare foundin the plan,it could mean that the previous plan selectionwas flawed and a sub-optimal plan might have been selected. We wish to usethe estimation errorsfound in the plan as an indicator for potentialsub-optimality in the plan.

Estimation errors, dependent upon where they occur, could have different impacts on the cost of the plans. There is probably no single or even a set ofbest threshold valuesfor all possibleoperations, queries, anddata distributions. In order not to complicate the discussion, here for simplicity, we assume a single threshold value for each query.We shall investigate the potential of using multiple thresholds in the future research.

It is important to find a good threshold value. If the threshold is set too high, sub-optimality in theplan can easily elude the checkpoints. On the other hand, if the threshold is set too low, minor errors that would not cause any change to the plan can trigger the gathering of statistics (i.e., a false alarm).

We propose to use a dynamically adjustable threshold value, described as follows. First, the threshold is assigned a low initial value e.g., 5%. When the estimation errors exceed the threshold, we gatherdesired statistics (according to the heuristics in Section 6.2.2) and generate a "new" plan. If the "new" plan is really a new one, we decrease the threshold value(because a lower threshold couldhave also led to the same change); otherwise, increase the threshold (as itmight have been set too low and caused the false alarm).

There are many possible ways to adjust the threshold value. One simple way is to increase or decrease it by a fixed amount (e.g., 5%).One can also change the threshold value by an amount proportional to the amount of errors, etc. We shall leave these options for future research.

In the following, we summarize the essence of the lazy re-optimization into the following algorithm.The variable “plan” stores the plan of the query andthe “statistics” stores whatever the sufficient statistics that have been gathered. The“cards” stores the actual cardinalities recorded at the checkpoints, while the “est_cards”stores the estimated cardinalities derived by

the optimizer using available statistics. "T" is the threshold and the"errors"stores the estimation errors at the checkpoints. "changed" is a Boolean flag indicating whether a differentnew plan has been generated by the optimizer or not. "re_optimize()" represents the process of generating a plan using the statistics available in the catalog and the accurate statistics gathered and stored in "statistics".

```
Algorithm Lazy-Re-Opt (plan, statistics, cards, est_cards,T)
{
 errors=compare(cards, est_cards);  /* estim.errors*/
 if(max(errors) >T){
   get_desired_stat(statistics, errors);  /* Sec. 6.2.2
   changed= re_optimize(plan, statistics);
 if (changed==true) /* Sec. 6.2.3
    increase(T);
 else
   decrease(T);
 }
}
```

Figure 6.2 Lazy Re-optimization

6.3 Adaptive Re-optimization

An optimal plan can degenerate to a sub-optimal one once the database has undergone substantial changes. To guarantee the optimality of the plan for the entire lifetime of the query, an adaptive scheme is devised. The adaptive re-optimization can be achieved by constantly monitoring the executions and performingnecessary re-optimizations. There are two simple ways to implement the adaptive re-optimization. The first possibility is to couple the cardinality monitoring process of the laze re-optimization with the eager re-optimization, that is, to gather all the sufficient statistics whenever estimation errors are found to be greater than the threshold. Another possibility is to simply extend the cardinality monitoring period of the lazy re-optimization to the entire lifetime of the query, and the lazy re-optimization immediately becomes an adaptive re-optimization scheme. By detecting and remedying the sub-optimality timely, the adaptive re-optimization may enable queries to run in the most efficient ways for the entire lifetime of the queries.

7. CONCLUSIONS

In this paper, we propose a comprehensive re-optimization framework for an important and large class of queries – repetitive queries. We first discussed statistics that are sufficient to find the best plan for a query. The proposed sufficient statistics make re-optimization of queries a realistic and achievable goal. Then, we discussed different ways to gather the sufficient statistics and presented two innovative methods to conduct re-optimization – the eager and the lazy re-optimization. The eager re-optimization attains the optimal plans in one step, while the lazy re-optimization in multiple steps. We have also designed an adaptive re-optimization method to adjust the plans dynamically so that the queries can always be executed in the optimal fashions for their entire lifetime. The approximate re-optimization presents an efficient and effective alternative to refining query plans.

In the future, we shall extend the coverage of queries to those with other useful operators and aggregate functions. Although we have proved that with the proposed sufficient statistics on hand, one can always derive the optimal plans (i.e., an eager re-optimization), we still need to verify by experiment how effective the proposed heuristics in the lazy approach are and examine the quality of the plans generated by the proposed approximate re-optimization method.

8. REFERENCES

[1] Babu, S., Bizarro, P., Dewitt, D., Proactive Re-optimization, *ACM SIGMOD Conf., 2005*, 107 -118.

[2] [Chaudhuri, S., Narasayya, V., Ramamurthy, R., Diagnosing Estimation Errors in Page Counts Using Execution Feedback, *ICDE, 2008*, 1013 -1022.

[3] S. Christodoulakis. Implications of certain assumptions in database performance evaluation.*ACM TODS*, 9(2): 163-186, June 1984.

[4] Ioannidis, Y., Kang, Y., Left-deep vs. bushy trees: an analysis of strategy spaces and its implications for query optimization, *ACM SIGMOD Conf., 1991*, 168 – 177.

[5] Kabra, N., Dewitt, D., Efficient Mid-Query Re-Optimization of Sub-Optimal Query Execution Plans, *ACM SIGMOD Conf., 1998*, 106-117.

[6] Markl V., Robust Query Processing through Progressive Optimization, *SIGMOD Conf, 2004*, 659-670.

[7] Muralikrishna M., DeWitt D.: Equi-Depth Histograms for Estimating Selectivity Factors for Multi Dimensional Queries. *SIGMOD Conf. ,1988*, 28-36

[8] Ng, K., Wang, Z., Muntz, R., Nittel, S., Dynamic Query Re-optimization, *SSDBM Conf, 1999*. 264-273.

[9] Piatetsky-Shapiro G., Connell C.: Accurate Estimation of the Number of Tuples Satisfying a Condition. *SIGMOD Conf. 1984*,256-276.

[10] Ramakrishnan, Gehrke, *Database Management Systems, 3rd Ed.*, McGraw Hill, 2003.

[11] P. Selinger , M. M. Astrahan , D. D. Chamberlin , R. A. Lorie , T. G. Price, Access path selection in a relational database management system, SIGMOD Conf., 1979, 23-34.

[12] Stillger, M., Lohman, G., Markl, V., Kandil, M., LEO - DB2's Learning Optimizer. *VLDB, 2001*.

[13] Yu, F., A Framework for Re-optimization of Repetitive Queries, http://www.cs.siu.edu/~fyu/reopt.pdf

Database Performance Monitoring and Tuning Using Intelligent Agent Assistants

Sherif Elfayoumy and Jigisha Patel
School of Computing, University of North Florida, Jacksonville, FL,USA

Abstract - *Fast databases are a necessity in today's E-commerce environment. Information from the database has to be provided at fast rates to impatient customers, otherwise they will move on to competitors with more attentive services. Corporations needing up-to-date internal information cannot wait for long running queries that process numbers and produce detailed statistics. Instead, they need databases capable of providing data in fractions of seconds to compete in today's economy. Database performance tuning has always been one of the most important tasks for database administrators (DBAs). This research presents an intelligent agent assistant to aid DBAs in performance monitoring tasks and the automation of resolution actions. The assistant notifies DBAs when performance problems are detected and resolved. DBAs are expected to provide the agent assistant with definitions for the performance problem conditions and the possible resolution actions. They must also develop a notification communication mechanism for the problems the assistant agent monitors.*

Keywords: Intelligent Agents; Performance Tuning; Database Performance.

1 Introduction

In today's competitive business environments, application performance plays a major role in business success. Bad application performance can result in unhappy customers and revenue loss. Usually applications have their data stored in large database repositories. Database performance affects applications' response times, which eventually affect the end user's experience. Database performance is directly affected by the Database Management Systems (DBMS) resources allocation where many resource dependencies are involved, which makes performance troubleshooting a complex task.

Highly paid and experienced Database Administrators (DBAs) are typically needed to perform the database performance-tuning task. This task has to be performed on a constant basis to cope with the changes in performance due to database growth and workload changes. Achieving peak database performance is a difficult task but is not impossible. Database tuning requires DBAs to identify resources not properly tuned and then tune them to obtain better performance.

The intelligent agent assistant is a knowledge-based system that perceives its environment, reasons to interpret perceptions, draws interfaces, solves problems, and determines Actions; and acts upon that environment to realize a set of goals or tasks for which it was designed. There are different types of intelligent agents. For example, user or personal agents take actions on behalf of the user to perform tasks such as gathering information on a subject of interest, preparing customized news snippets, setting up and prioritizing e-mails according to the user's preference, and playing games. Another type of intelligent agent is the monitoring and surveillance agent, which can observe and report on equipment, devices, and systems. This type of agent can automate tasks like tracking equipment inventory, planning, and scheduling order deliveries.

Intelligent agent assistant (IAA) can be employed to tune the performance of DBMSs on behalf of DBAs, which should substantially reduce the cost of database ownership. IAA will continuously provide effective monitoring and allow the DBMS to respond quickly to performance issues. It should also continue to improve the DBMS performance for both dynamic and static workloads. It also should be platform independent, easy to maintain, scalable, extensible, and able to communicate.

2 Background

Many efforts have already been made towards automating the control of DBMS resources. An automated system should be able to identify and readjust DBMS resources. Anolan et al., proposed a design and implementation of a global self-tuning architecture [1]. Their research considered performance tuning as a global issue, given changes of a single parameter can affect the performance of other operations. It described the proposed architecture and also discussed system's integration within PostgreSQL in an implementation based on software agents. Software agents used to provide a more flexible approach to include automatic tuning features in PostgresQL DBMS.

JP Bigus et al., described AutoTune, an agent-based approach to automated tuning [2]. This agent-based approach to automated tuning did not require prior knowledge of the controlled system being tuned. AutoTune enabled target systems to expose workload metrics (e.g., RPC arrival rates), configuration (e.g., processor speeds), and service levels (e.g., response times), as well as means to manipulate tuning controls (e.g., admission control parameters). This agent-based approach constructs a generic model of the target system (e.g., by training a neural network) and from this derive a controller. A prototype of AutoTune agent was

implemented in the Agent Building and Learning Environment (ABLE)[2]. Also, Darcy G. Benoit demonstrated the automatic diagnosis of performance problems in DBMSs [3]. This dissertation discussed and demonstrated ideas for diagnosing database performance problems at small subset levels. The diagnosis model was tested on workload changes and database size changes. The model creates diagnosis trees for different types of databases and their workloads.

Chung et al., focused on goal-oriented buffer pool management [4]. In this paper, a performance index was calculated based on the response time of the buffer pool. They presented a goal-oriented approach in which the user can specify a buffer pool response time goal and total count for all buffer pools in the database. This goal-oriented approach dynamically changes each buffer pool size, while maintaining the total number of buffers. This dynamic tuning results in a great improvement in database performance. Also, Agrawal et al., investigated whether the selection of materialized views and indexes can be automated for DBMSs [6]. Database performance can be improved with a correct combination of indexes and materialized views. They developed alogrithms and an architecture; and, explained how they were able to identify a small set of materialized views and indexes, despite a totally different structure.

Chaudhuri and Narasayya explained database technology as RISC-style data managers with self tuning capabilities [5]. Each individual component is self tuning and exhibits predictable performance by eliminating the need for manual tuning by a DBA. These components are capable of determining which database statistics are essential for boosting performance and make sure they are updated frequently. They explained that database tuning must consider the relationship between workload characteristics, knob settings, and the resulting performance, in a quantitative manner.

Oracle provides a database "grid control agent," which can help DBAs monitor and maintain Oracle databases. Oracle's agent provides a good mechanism for database administration tasks and sends warning/critical notices, depending on setup. Jobs can be setup as shown in Figure 1 by providing values for job name, description, and SQL script fields. User should also select one or more databases from the list on which this job will run.

The desired schedule should be setup by clicking the schedule tab, as shown in Figure 2. Oracle's agent provides several scheduling options such as one time only, monthly, yearly, and user-defined intervals. It also has options for allowing simple SQL statements or using complex SQL script stored in the database server for database monitoring. Despite these good features, it is missing some important features, such as a mechanism to take automatic actions and an option to add multiple SQL statements.

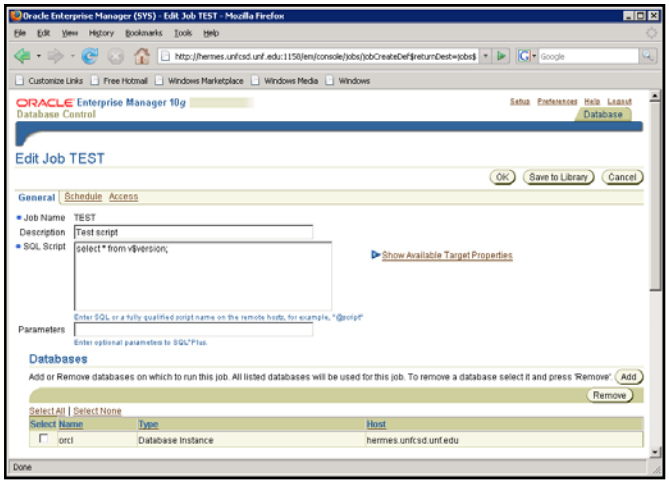


Figure 1: Create Job (Oracle)

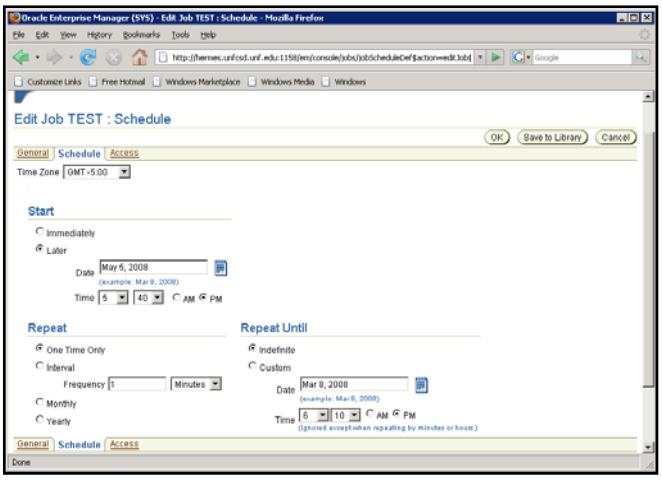


Figure 2: Schedule Job (Oracle)

Microsoft SQL server also provides an agent tool to help DBAs monitor SQL server databases. New jobs can be scheduled by providing job name, schedule type, start time, and frequency to run the job, as shown in Figure 3. Job status notifications can be sent using email, pagers, or net send commands by providing email address, pager number or machine address. This agent provides several options for scheduling and notification, but is not flexible enough to support non-Microsoft SQL server databases.

A review of the published literature on automated performance tuning approaches and tools provided no evidence of the availability of a generic tool to assist DBAs in managing a variety of DBMSs in a heterogeneous environment. In this paper we present an intelligent agent assistant that was developed to work for a variety of databases such as Oracle, SQL Server, MySQL, to send SMS notifications to DBAs, to be easily manageable for most relational databases through the use of XML scripts, to use standard SQL queries to gather performance information from databases, rather than platform dependent queries such as T-SQL and PL-SQL, and to allow DBAs to provide any number of performance tuning queries.

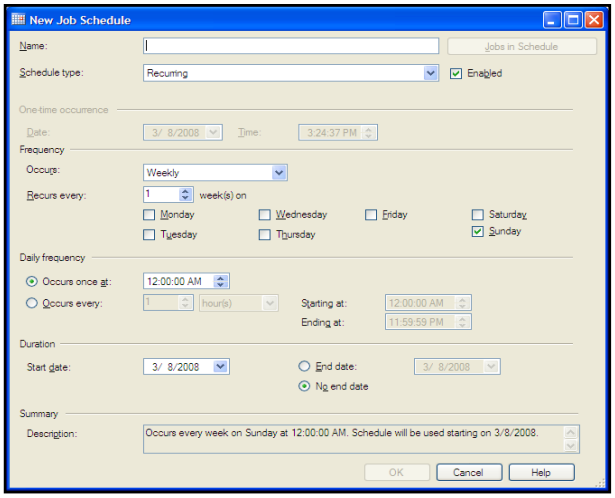


Figure 3: New Job Schedule (SQL Server)

3 Intelligent Agent Assistant

In this research effort an intelligent agent assistant was designed and developed to operate on user defined rules. In the form of database performance tuning rules, SQL queries are used to measure certain performance metrics of the database server and/or to modify some of the performance parameters, if need be. Once the agent detects unsatisfactory performance measures, it performs the actions specified in its rule set, so it can rectify the performance issues. DBA needs to define IAA rules by specifying resources to monitor, queries to collect performance metrics, threshold levels of minimum acceptable values for performance metrics, corrective actions, if performance problems are identified, and frequency of performing these checks.

The IAA can run continuously in the background without any human intervention. The system was constructed to use multithreading, specifically Java threads. Multithreading provides for more responsive GUI, leveraging multiprocessing hardware, simpler modeling, and asynchronous processing of rules. Once the agent assistant has started, it starts the first thread, which is the main/parent thread. The main thread reads stored rules information from the knowledgebase, implemented in XML, and spawns off other child threads, according to a specified schedule, to monitor resources and perform corrective actions. The main thread keeps running until stopped through the user interface program (GUI). DBAs can add, edit, or delete rules through the GUI. Additions and changes are stored in the XML-based knowledgebase and take effect only after restarting the agent.

IAA uses an XML file to store all the information required to connect to the DBMS, including server name, instance name, user id, encrypted password, queries required to gather resource information, frequency at which resource information is gathered, formulas required to determine if performance problems exist, resolution actions, and phone number to notify when corrective actions are performed. XML was chosen to store agent's rules because of its extensibility and ease of use to communicate between different system components, which otherwise would be unable to communicate or would be rather tightly coupled. Figure 4 shows IAA architecture and interaction between GUI, XML file, and agent. The XML file is validated against the DTD schema file exhibited in Figure 5.

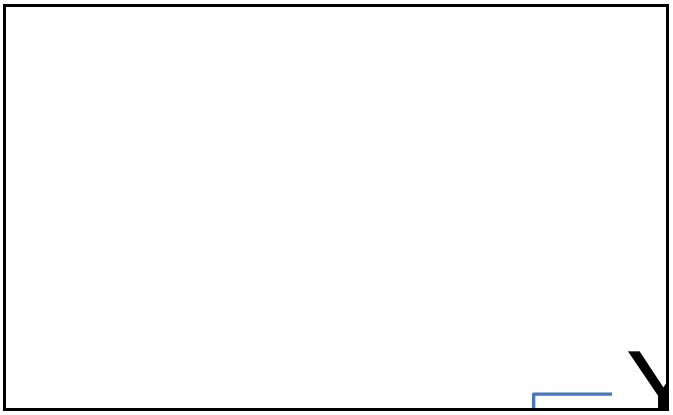


Figure 4: IAA System Architecture

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE document [
<!ELEMENT document (database)*>
<!ELEMENT database (connstring+,userid+,passwd+,rule*)>
<!ATTLIST database
          name CDATA #REQUIRED>
<!ELEMENT connstring (#PCDATA)>
<!ELEMENT userid (#PCDATA)>
<!ELEMENT passwd (#PCDATA)>
<!ELEMENT rule (query*,interval+,condition+,action*,min?,max?,increment?
,decrement?,dbbounce?,dbanotify?)>
<!ATTLIST rule
          name CDATA #REQUIRED>
<!ELEMENT query (#PCDATA)>
<!ATTLIST query
          name CDATA #REQUIRED>
<!ELEMENT interval (#PCDATA)>
<!ELEMENT condition (#PCDATA)>
<!ELEMENT action (#PCDATA)>
<!ATTLIST action
          name CDATA #REQUIRED>
<!ELEMENT min (#PCDATA)>
<!ELEMENT max (#PCDATA)>
<!ELEMENT increment (#PCDATA)>
<!ELEMENT decrement (#PCDATA)>
<!ELEMENT dbbounce (#PCDATA)>
<!ELEMENT dbanotify (#PCDATA)>
]>
```

Figure 5: Validation Schema

Figure 6 shows details of a database node in XML. It has a <connstring> element, which contains the connection string for the database. The <userid> element contains the database administrator's user name and <passwd> contains the password.

The rule node has a number of <query> elements; each is identified by a name parameter (Figure 7). These queries are used to gather database performance statistics.

Part of the XML rule node is shown in figure 8. The <interval> element is another node that stores the execution frequency of a rule, in minutes. The <condition> node contains a formula prepared by the DBA to determine if a performance problem exists. The condition may use some of the values returned by the queries, where each query is

identified by name. A condition can be a combination of queries with logical operators (AND and OR). If a condition is satisfied, the associated actions will be performed, as prescribed in the <action> element. The DBA can also specify multiple actions.

```
<database name="oracle">
    <connstring>
        jdbc:oracle:thin:@hermes.ccec.unf.edu:1521:ORCL
    </connstring>
    <userid>
        system
    </userid>
    <passwd>
        unf
    </passwd>
```

Figure 6: XML Document's Database Node

```
<rule name="db_cache_size">
    <query name="1">
        select value from v$sysstat where name in ('db block gets')
    </query>
    <query name="2">
        select value from v$sysstat where name in ('consistent gets')
    </query>
    <query name="3">
        select value from v$sysstat where name in ('physical reads')
    </query>
    <query name="4">
        select name from v$database
    </query>
    <query name="5">
        select value from v$parameter where name like '%db_cache_size%'
    </query>
```

Figure 7: XML Document's Query Element

```
    <interval>
        3
    </interval>
    <condition>
        (1 - ( Q1 / ( Q2 + Q3 ))) * 100 &lt; 90
    </condition>
    <action name="1">
        alter system set db_cache_size = Q5
    </action>
    <min>
        0
    </min>
    <max>
        251658240
    </max>
    <increment>
        1
    </increment>
    <decrement>
        0
    </decrement>
    <dbbounce>
        N
    </dbbounce>
    <dbanotify>
        904-236-1981
    </dbanotify>
</rule>
```

Figure 8: XML Document's Rule Node

The <min> and <max> nodes are used for providing acceptable limits for a given resource. The elements <increment> and <decrement> are used to store values representing the amount of increase or decrease required to adjust a database resource when needed. The node <dbbounce> is a flag to indicate whether the database needs to be bounced, after an action is performed, for the change to take effect. The expected value of that node is 'Y' for yes or 'N' for no. Finally, if element <dbanotify> has any value other than "0", then an SMS notification with the rule details will be sent to the DBA, if the associated action is performed.

3.1 Use Case: Tuning Buffer Cache Hit Ratio

A use case was developed to demonstrate IAA's operation and the ease of use to customize the performance rules.
The Buffer Cache Hit Ratio is an Oracle metric used for the rate at which Oracle finds data blocks in main memory. A correctly tuned buffer cache can significantly improve overall database performance. The traditional trade-off between efficiency and performance exists here. A small buffer cache would result in more disks I/O operations, which reduces performance. On the other hand, a big buffer cache is a waste of the server's scarce memory (less efficient). The hit ratio depends on physical and logical reads. A logical read occurs whenever the user requests data from the database, which is in memory or server hard disk. A physical read occurs only when the data must be read from hard disk. Oracle stores physical and logical read details in the V$SYSSTAT table. The hit ratio value can be determined with the following SQL query:

SELECT name, value FROM v$sysstat
WHERE name in ('db block gets', 'consistent gets', 'physical reads');

The cache hit ratio is calculated using the following formula:

Hit ratio = 1 - (PR / (DG + CG))
Where: PR = Physical reads,
DG = Db block gets, and
CG = Consistent gets

Ideally, the cache hit ratio should be greater than 90%. If it goes below the desired value, the database parameter "DB_CACHE_SIZE" Will need to be increased. To create an agent rule, for tuning based on the buffer cache hit ratio, SQL statements should be converted into single column queries named Q1, Q2, etc. The condition and action can use these query names. The agent rule created for this use case is shown in Figure 9.

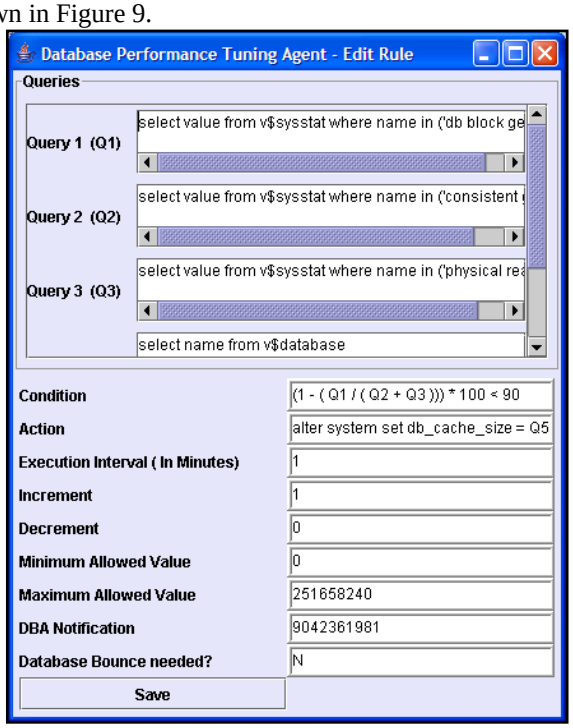


Figure 9: Use Case - Tuning Buffer Cache Hit Ratio

4 Conclusion

DBMS resources must be carefully allocated to achieve a good balance of performance versus efficiency. This task is complicated by the fluidity of DBMSs. Data and workloads keep changing and resource allocation must cope with these changes. The only solution to this performance dilemma is to keep monitoring the database continuously. The automation of some database monitoring and resource management aspects lessens the burden on expensive DBAs. The use of intelligent agent assistants is an effective approach to solving the database-tuning problem. We plan to continue testing the IAA using more use cases and enhance its operation by creating more flexible scheduling, create a web interface to manage rules, allow rule dependency, and allow for other types of notifications such as email.

5 References

[1] Anolan Yamile Milanes and S. Lifschitz. "Design and Implementation of a Global Self-tuning Architecture"; 20th Brazilian Symposium on Databases, pp. 7-11, 2005.

[2] Bigus J.P., J.L. Hellerstein, T.S. Jayram, and M.S. Squillante. "AutoTune: A Generic Agent for Automated Performance Tuning"; IBM Thomas J. Watson Research Center, pp. 4-7, 2000.

[3] Benoit G. Darcy. "Automatic Diagnosis of Performance Problems in Database Management Systems"; Proceedings of the Second International Conference on Automatic Computing, pp. 70-77, 2003.

[4] Chung Jen-Yao, D. Ferguson, G. Wang, C. Nikolaou and J. Teng. "Goal Oriented Dynamic Buffer Pool Management For Data Base Systems"; Proceedings of the 1st International Conference on Engineering of Complex Computer Systems, pp. 7-10, 1995.

[5] Chaudhuri Surajit and G. Weikum. "Rethinking Database System Architecture: Towards a Self-tuning RISC-style Database System"; IBM Thomas J. Watson Research Center, pp. 5-8, 2000.

[6] Agrawal Sanjay, S. Chaudhuri, and V. Narasayya. "Automated Selection of Materialized Views and Indexes for SQL Databases"; Proceedings of the 26th International Conference on Very Large Databases, pp. 496-505, 2000.

Parallel Data Loader and OLAP Operations

Sherif Elfayoumy and Nishant Patel
School of Computing, University of North Florida, Jacksonville, FL, USA

Abstract - *Constructing fact tables and performing OLAP operations are some of the most essential but expensive operations in data warehousing. OLAP operations require aggregation on many combinations of each dimension attribute. As the number of dimensions increase, it becomes very expensive to compute summary data cubes because the cost of required computations grows exponentially with the increase of number of dimensions. Query processing for these applications requires different views of data for analysis and effective decision-making. As data warehouses grow, parallel processing techniques can be applied to enable the use of larger data sets and reduce the time for analysis, thereby enabling evaluation of many more options for decision-making. Architecture for computing OLAP operations in parallel, using multiple processors is presented in this paper.*

Keywords: Parallel Computing; Big Data; OLAP; Data Warehousing;

1 Introduction

The challenges of maintaining response times of data warehouses as they grow larger and more complex have given rise to new technology solutions like OLAP, and multi-dimensional databases. These solutions are attractive since they ensure the integration of data and give users a refined approach to decision-making. Online Analytical Processing (OLAP) is a database acceleration technique used for deductive analysis. The main objective of OLAP is to have constant-time or near constant-time answers for the many typical queries. For example, in a database containing salesmen's performance data, one may want to compute, online, the amount of sales done in a given city for the last 10 days, including only salesmen who have 2 or more years of experience. Using a relational database containing sales information, such a computation may be expensive. Using OLAP, however, the computation can typically be done online. To achieve such acceleration one can create a cube of data, a map from all attribute values to a given measure. In the example above, one could map tuples containing days, experience of the salesmen, and locations to the corresponding amount of sales. The biggest challenge in OLAP now is the sheer volume of data which needs to be handled. Large organizations are usually awash with data, and OLAP techniques are being used to try to make sense of these vast data volumes. However, most OLAP systems impose limits on data volumes, and cannot scale to the level now demanded of them.

Creating a common fact table containing information from all the data sources and issuing queries against that large fact table requires a considerable amount of time. It takes a significant amount of time to perform any of the basic OLAP operations such as roll-up, drill down, pivot, slice and dice on a big fact table. Performing these operations in parallel by dividing the work among all the available processors (nodes) should reduce the overall response time. As shown in figure 1, data from the operational databases are fetched in parallel by multiple processors, and each processor creates a fact table locally. Once the creation of the fact table is completed, the user can start querying that fact table using the available OLAP tool. All the processors compute the results in parallel and then the manager combines results from all the processors and displays the result in the form of a grid or a chart.

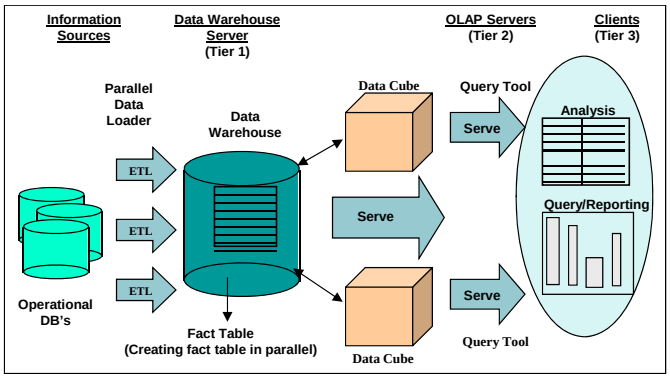


Figure 1: Data Loader and OLAP Tool

1.1 Bottom-up Approach

The Initial approach to parallelism that we considered was based on the Bottom-Up Cubing approach. Bottom-up approaches reuse previously computed sort orders and generate more detailed group-by result sets from less detailed ones. Bottom-up data cube construction methods calculate the group-by result sets in an order that emphasizes the reuse of previously computed sorts. All OLAP operations (Roll Up, Drill Down, Pivot and Slice & Dice) can be done using the bottom-up approach. For example, to do a Roll Up operation on a particular level, the only procedure required by this method is the aggregation level. By passing an aggregation level, this method will distribute cube dimensions to all the computing processors [1][2]. The main logic behind this bottom up sorting is that if it is known that values of attribute A are already sorted, then there is no need for the complete resorting of AB. The total computation time of the bottom up method is dominated by the number of such single attribute

sorts. However, after running some experiments it proved that the communication cost of this approach was much higher than the new approach.

Figure 2 provides an illustration of the recursive subdivision of a four dimensional space. In this case, BUC (Bottom-Up Computation) will first partition (i.e., sort) the input set on A. It will then aggregate on <*a*1> before partitioning <*a*1> into its <*a*1*b*1> and <*a*1*b*2> components. This recursive partitioning will continue until the last dimension has been reached. Eventually, the backtracking will return the algorithm to the <*a*2> partition, at which point the whole process is repeated.

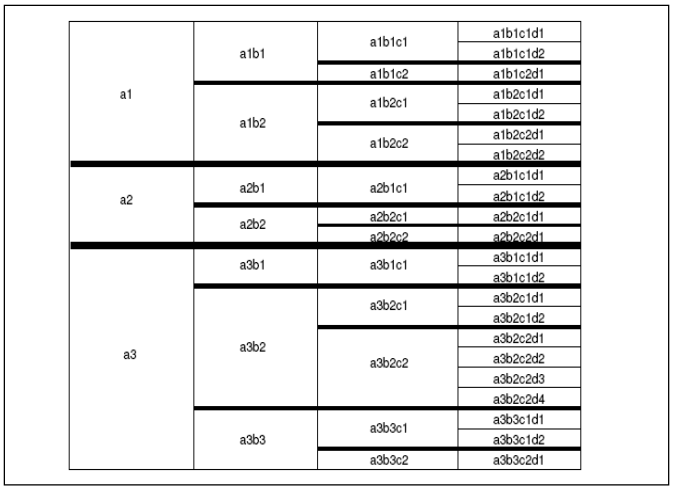


Figure 2: The bottom up "perspective"

The BUC algorithm (Figure 3) is well suited to sparse, high dimension data cube problems for these reasons.

- In step 2 of the algorithm, it checks to see if the size of the current partition is equal to one. If it is, then there is no value in continuing the recursion since no further partitioning can be performed. We therefore write out the aggregates for all ancestors and return immediately. For example, when it encounters the tuple <*a*1*b*1*c*2>, it knows it can write the aggregate value for <*a*1*b*1*c*2*d*> without further processing. Because many partitions will in fact have a size of one in sparse spaces, this *short circuiting* can significantly improve performance.
- As the algorithm recursively partitions the input set, BUC divides the data into smaller and smaller segments. Consequently, it is increasingly likely that these partitions fit entirely into main memory, possibly reducing the reliance on more expensive external memory sorting.

```
Input: The partition to be aggregated,
plus the current dimension d.
Output: A single record that represents
the aggregated input.
1: Aggregate input relation
2: if input count == 1 then
3: Write ancestor records and return
4: end if
5: Write output record from Step 1
{/* process remaining partitions */}
6: numDims = total number of dimensions
in data cube
7: for dim = d; dim < numDims; dim++ do
8: c = cardinality of dimension dim
9: Partition input on its c unique values
10: for each of the c partitions in the
input set do
11: Recursively call BUC(partition,
dim+1) using the current partition as
input
12: end for
13: end for
```

Figure 3: Bottom-Up Algorithm

As noted above, the order in which the attributes are partitioned is also important. Specifically, BUC partitions the attributes in order of decreasing cardinality. In so doing, it minimizes the use of large sorts since the maximum cardinality dimension will immediately split the data into as many small partitions as possible. Experimental results for BUC demonstrate approximately a factor of two performance advantage with respect to alternative data cube algorithms in sparse spaces. However, the benefit of BUC tends to be limited to these full cube high dimension problems. When denser views - the kind typically found in the lower levels of the lattice - are required, the time to recursively partition input sets could dominate the run time since very little short-circuiting takes place. As a result, BUC is ill suited to problems in which large sparse views either not exist or are not required.

The Bottom-Up approach requires reuse of previously computed cubes, which is expected to reduce the number of required computations [3]. However, a small experiment using the Bottom-Up approach showed that there is a lot of overhead in reusing a computed cube. Because each worker node starts with its own computed cube, sends that cube to the other workers, and other workers send computed cubes to that worker, the time required to send cubes was greater than the time required to compute cubes locally. Thus, this project doesn't use a data-local approach where cubes are computed locally instead of sending them back and forth. Another problem of using the Bottom-Up approach is that it is recursive [3]. This recursive nature turned a small experiment into a complex job. The excessive use of memory made this approach unfeasible using desktop nodes. Balancing the load between workers is also a problem. In some queries, one worker could do all the computation while other worker nodes remain idle. Finally, the data size at worker machines increases tremendously when a single worker machine tries to compute the whole tree.

2 Parallel Data Loader and OLAP Operations

The appeal of parallel processing is especially strong for data warehouse environments, due to the inherent nature of these environments. OLAP systems emphasize interactive

processing of complex queries. Given this requirement, as well as the often extreme size of warehouses, methods are clearly needed for more rapid query execution. By partitioning data among a set of processors, OLAP queries can be executed concurrently, in parallel, potentially achieving linear speedup and thus significantly improving query response time [4][5][6]. In addition to these OLAP queries, another problem data warehouses face is the creation of fact tables. Using a large amount of data and large number of data tables increases the time required for creating the fact table. In order to solve this problem, new approaches for creating fact tables need to be explored. This chapter discusses a new approach that proved to address these efficiency issues.

2.1 Data Loader Design

The main function of the data loader portion of this research was to connect to remote databases and get the appropriate data to create a local fact table. The data loader is also independent of the platform in the sense that the remote database can be implemented on any RDBMS. Thus, one of the main requirements is to have Internet connectivity at the remote sites and to have a DBMS server running on each client's machine. The following section discusses the available modes for storing data cubes.

Storage Modes: Once data is fetched from multiple sources, there must be a way to store it into a centralized location. There are three main choices for storing data. These are Multidimensional OLAP, Relational OLAP and Hybrid OLAP [3][7]. ROLAP was used to take advantage of combining transactional data and historical data stored in a data warehouse. Processing ROLAP can be very slow on single processor systems, but it can be improved greatly by utilizing parallel processors. The design of the data warehouse database schema should incorporate the principles of dimensional modeling so the dimension tables and fact tables represent the business data and the way clients will view and query the data. Most dimensional modeling results in a star or snowflake database schema. Such schemas facilitate cube design and reduce the number of multiple-table joins when querying the database to process dimensions and cubes [1]. This project uses a star schema as the database schema.

Creating the Fact Table: Initially all workers wait for a request from the manager. The manager broadcasts a message to all workers that the user wants to create the fact table. Each worker gets the address from where it has to fetch the data. Workers will connect to the remote databases and start retrieving data. Depending on the criteria selected by the user to create the fact table, the workers perform queries on their remote databases and fetch the needed rows. This approach assumes all the dimension tables needed to build the fact table are already stored locally in each worker machine, as illustrated in figure 4. This makes it unnecessary to transfer dimension tables from the manager machine to the worker machines. In summary, worker machines need to fetch the transaction and historical data from the remote databases. Using the dimension tables and the transaction table, worker machines create fact tables locally. Once the creation of the fact table for one remote database is complete, the worker checks whether there are any more addresses available for which data needs to be fetched. If there are not any addresses available, it sends a notification message to the manager indicating that it is done with fact table creation. If there is some address available, the worker fetches data from that address and appends rows to the already existing fact table.

The initial approach investigated was to send the fact table data from the workers to the manager. However, after conducting further research and collecting results from a small experiment it was clear that since the manager sends the OLAP queries back to the workers, it is better to keep data local to the worker nodes. The biggest advantage of using this approach is that a large space on the manager computer is not needed, because the fact table is divided into several sub fact tables on worker machines. The second advantage is that only minimal communication is needed between the manager and workers; the manager does not send data to workers but rather sends processing requests.

2.2 OLAP Design

Figure 5 shows that the manager sends OLAP queries to all the worker machines. Each worker then computes the results locally and sends them back to the manager. Finally, the manager combines the results collected from the different workers and presents them to the user in the form of a table or a chart. Since each worker has part of the fact table and the dimension tables locally, communication cost between manager and worker is minimized. In addition, there is no need for inter-worker communication where workers need to communicate only with the manager. Each worker has the same copy of dimension tables, which are typically not very big. It is also possible that the manager makes sure each worker has the latest copy of dimension tables. Thus, whenever the manager sends an OLAP query to the workers, they do not need to fetch the data from the manager, but only compute parts of the fact table they store locally and send the results back to the manager. Since the number of workers available divides the fact table and each worker has to manage only its part, queries are processed faster when compared with a single node.

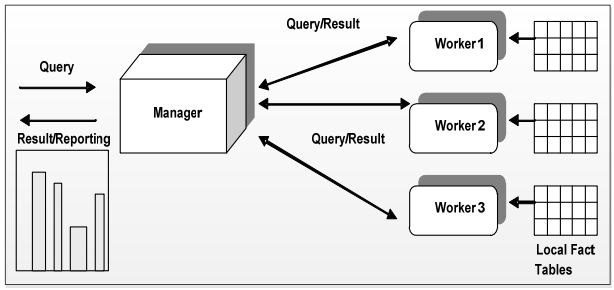


Figure 5: Parallel OLAP Architecture

2.3 Data Loader Implementation

User can select the criteria for building the fact table. He can select the dimension tables and the fields from the remote table he wants to select as measures for that fact table. Once user has selected dimension tables, primary keys for those tables are automatically imported in the fact table. These primary keys become foreign keys in the fact table. User can also select measure fields and the type of operation he wants to perform on them. Once the creation of the fact table is completed, workers start populating the fact table with values. All the information regarding the remote tables, such as connection string and remote table name are stored in the remote URL table. Once the manager sends a message to the workers to start working on the creation of the fact table, each worker fetches one connection string from this remote URL table. Once a worker is done with creating fact table data for that connection string, it asks the manager whether there is any more data to be fetched. For example, if there are four remote databases and if there are two worker machines, the first worker will fetch data from the first remote database and the second worker will fetch data from the second database. The worker that finishes earliest will start fetching data from the third database. Once a worker is done with the creation of the fact table, it sends a message to the manager that it is done. Once the manager receives acknowledgement from all the workers it notifies the user that the creation of the fact table is completed.

2.4 OLAP Implementation

Initially users have to select the attributes from the dimension tables to perform queries, as shown in figure 6. Users are also allowed to select the measure on which they want to perform the query, so the user can control the way the query is performed.

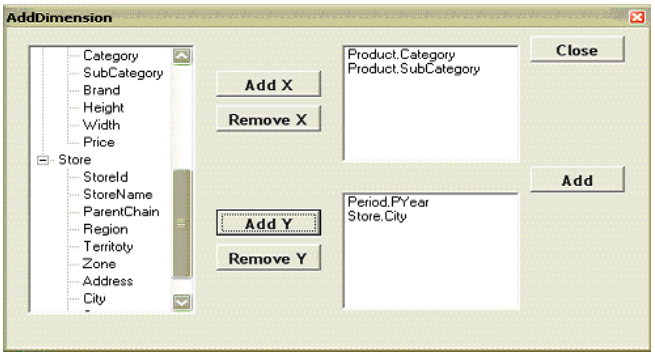


Figure 6: OLAP Selecting Attributes

After selecting attributes, user can select the value of those attributes; for example, if he selects an attribute named "Year" from the Period table, he can select which year he wants to look into, say 2003, for example, as illustrated in figure 7. According to this all OLAP operations become generic, i.e. there is no need for customized implementation of the different OLAP operations. Once the user has selected all the fields he wants to see, the manager will send all requests to all available worker nodes. Each worker gets this information and starts working on computing the results. Each worker node uses its dimension tables and the locally stored part of the fact table to compute the results. Once it computes the results of all the queries, it sends them back to the manager. The manager waits for a response from all workers. Once all workers send their results to the manager, it combines the results and presents them in the form of a table or a chart as seen in figure 7.

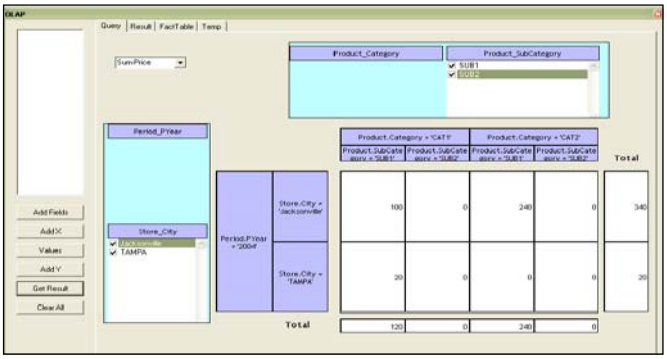


Figure 7: OLAP Performing Query

Users also have the option to select the measure on which to perform the analysis. In addition, users can select to view the results in the form of bar or pie charts (Figure 8), as well as the default tabular format. These charts are implemented using .Net Charting graphics library.

Figure 8: OLAP Output of Query

3 Conclusion

The communication overhead of the Bottom Up approach increases as the size of data and the number of processors increase to the point that makes it not efficient for real applications. The Bottom Up approach reuses the cube computed by one worker node to compute other cubes by other worker nodes, which involves a large number of big communication messages between worker nodes. This paper presents a more efficient alternative that processes sub-fact tables on worker nodes locally eliminating the inter-node communication overhead. Though the new approach is more

efficient and scalable than its predecessors, it does not address important issues such as fault tolerance. If one of the worker nodes goes down, the partition of the data set allocated to those nodes (sub-fact table) will be lost. Since no other workers have access to, or retain a copy, the absence of that data might lead to an incomplete fact table. One solution to this problem is to maintain a record of the work being done by all worker nodes and the status of each node (redundancy). One viable option to implement this solution is to designate a node primarily for managing node failures (failure manager). This node can reassign work from a failed node to other active worker nodes. Extensive experimentation is needed to assess the efficiency of this solution and its impact on the regular computation.

4 References

[1] S. Muto and M. Kitsuregawa, "A dynamic load balancing strategy for parallel data cube computation," Proceedings of the 2nd ACM international workshop on Data warehousing and OLAP (November 1999), pp. 67–72.

[2] F. Dehne, T. Eavis, S. Hambrusch, and A. Rau-Chaplin, "Parallelizing the Data Cube," Distributed and Parallel Databases, 11, 2 (March 2002), pp.181– 201.

[3] F. Dehne, T. Eavis, and A. Rau-Chaplin, "Top-Down Computation of Partial ROLAP
Data Cubes," Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) 8, 8 (January 2004), pp. 223.

[4] S. Goil and A. Choudhary, "A Parallel Scalable Infrastructure for OLAP and Data Mining," Proceedings of the 1999 International Symposium on Database Engineering & Applications (August 1999), pp. 178.

[5] D. Pedersen and T. Pedersen, "Achieving adaptively for OLAP-XML federations**,"** Proceedings of the 6th ACM international workshop on Data warehousing and OLAP (November 2004), pp. 25 – 32.

[6] J. Stephens and M. Poess, "MUDD: a multi-dimensional data generator," ACM SIGSOFT Software Engineering Notes, Proceedings of the 4th international workshop on Software and performance, 2004.

[7] X. Zhang, L. Ding, and E. Rundensteiner, "Parallel multisource view maintenance," The International Journal on Very Large Databases, Vol. 13, No. 1, 2004.

Problems of Information Security in Temporal Databases: An Analytical Study

Ahmad Ali Al-Zubi
Computer Science Department, King Saud University, Riyadh, Saudi Arabia

Abstract: *Present paper is devoted to information security issues in temporal databases. All features and characteristics of temporal systems are considered and studied in this paper: main important types of temporal relations; as well as the principles providingconfidentiality, integrity, and reliability - the three basic components of information security in databases. The presented examples in this paper are made in relation to the problems of analytical informationsystems of the railway.*

Keywords: temporal database, relational database, information security, confidentiality, integrity, reliability, availability

1. Introduction

As the information society grows, the control systems limiting the people's access to information resources are becoming increasingly important. The process of improving the security of information resources always went along with the development of information technology and equipment. This development has allowed the information resources to improve themselves, giving to data some of the characteristics that were previously unaffordable luxury, such as dynamics, almost unlimited storage period and capacity.

The modern transportation systems are a perfect illustration of the systems with a high dynamics of objects. As an example, we can take the timetable of a passenger rail system. The need to preserve the history of changes of objects statuses in such systems requires a development of temporal databases [1,3]. The development of such systems is a nontrivial task, associated with the complex organization of data storage and processing. For example, a schedule of passenger trains is inherent in multiversion. Multiversion is associated not only with seasonal schedules, but also with a variety of ongoing changes in the case of repair works, holidays, and other, often unpredictable corrections in the movement of trains.

Note that, Temporal Databases (TDB), in addition to providing flexible work with historical data of operational systems, they also play an important role in analytical systems, since the accumulated historical data are of great value for planning and forecasting activities for the enterprises to identify trends and make decisions.

Among the tasks that must first be addressed in TDB, we should mention the problem of ensuring security of temporal data, which largely determine the quality of the functioning of established information-analytical systems. Despite a number of specific features and the importance of this problem in the current literature, it is not paid enough attention.

These circumstances explain the relevance of these studies to ensure the security of TDB. It is well known that the main factors that reduce the data quality of corporate information systems are not intentional attack of external foreign intruders, but hardware errors, administrators, applications and users [7]. For TDB, the semantics of which is more complicated than regular Relational Database (RDB), the truth of this statement is even brighter. Therefore, in this paper the author focuses on the protection of "historical" data from registered users. The problems of unauthenticated access, hacking, network security are not in the scope of this article and are not subject to the proposed research.

2. Basic Principles and Features Temporal Database Development

Until now, TDB did not yet received wide application. Deterrent to the technology development of TDB was a high requirement for resources to support multiversion of storage facilities. But due to the rapid development of technology, this factor has to recede to the second place. The cost of information in our time may greatly exceed the cost of equipment used itself. Let us formulate the basic tasks of information systems that require the use of temporal technology:

- Organization of storage and data access with temporal characteristics (eg, schedule of passenger transport);
- The need to store information that is no longer relevant over time, but awareness of this information is necessary in any moments of the past;
- Automatically saving earlier states of the object in order to log changes and access them if necessary;
- Storage and processing of streaming data.

In the last decade in Databases Development has occurred a separate scientific research trend in the field of temporal databases, including issues of data modeling, data organization in external memory, query languages, etc. Principles of development of the TDB implies that the database (DB) has to store all versions of objects and the user should be able to access all previous versions. Temporary expansion of the Database Management System (DBMS) implementing these

principles, automatically keepsa number of time-dependent versions of data objects.

The main concept in the temporal model is the object life period (lifespan). It describes a time interval associated with the existence of an object in particular state. Generally in contrast to the database, TDB has a number of specific features; their organization requires special technology - technology of temporal databases.

Traditional databases store a snapshot of the domain model. Any change of an object at time t leads to the inaccessibility of the state of this object in the previous point in time. There is a so-called "wiping" of data.

The main principles of the TDB technology are: storage of all states of the object since its creation to destruction, and providing access to any of the states at any given time. In other words, the main thesis of temporal systems is that for any data object that was created at time t1 and destroyed at time t2, all of its states in the time interval [t1, t2] are stored in the database and accessible to users. For efficient and safe operation of the TDB, its specific storage and data access should be taken into account in the security system.

Despite the merits and relevance of temporal technology and the existence of a separate scientific research field and development in the field of TDB, there are still problems in industrial realization of TDB. To some extent these problems are related to the choice of platform for TDB. The choice of platform depends not only on requirements of performance efficiency and functionality of the database, but also a number of other factors.

Research and prototyping of Temporal DBMS (TDBMS) are usually performed on the basis of a Relational DBMS (RDBMS). At the same time TDB is built on a relational system. Of course, this is not the best way to implement in terms of efficiency, but it is simple and accessible for systems developers and their users.

As a result, today most of the TDB is implemented in a relational database environment, although there are specialized temporal databases [3]. This situation is due to several factors that hamper the use of TDB in large enterprises. The main disadvantages of TDB to date are:

• Lack of a unified standard;
• Intolerance of developed tasks in the frame TDBMS;
• Noticeable lag in the power and reliability comparing with industrial relational systems;
• Weak support for the manufacturer.

However, for the benefit of RDBMS are the following facts:

• They are tested and proven over several decades;
• RDBMS vendors are constantly developing their products;
• Support the generally accepted standards;
• Relational media still the best for storage and data integrity;
• Most companies have many years of successful experience working with Relational Systems;
• The most of existing application systems developed by relational technology.

Speaking about the problems of using Relational Databases when dealing with temporal data, we can first note the following:

• RDBMS do not have the specialized tools for the development of TDB, whereas TDBMS is created for this task;
• RDBMS has no knowledge about the semantics of temporal relationship field and cannot control correctness of its values;
• Query languages of RDBMS are not designed to deal time.

Weighing all the "pros" and "against", enterprises, as a rule, sacrifice with built-in temporal functionality in favor of the reliability of the system, made investments, accumulated experience and prospects. Therefore, many new systems continue to be implemented in a relational environment.

Despite the absence of standard tools for developing Temporal Systems in RDBMS, their implementation on the basis of most relational systems is possible, but it requires in-depth design study of data model, knowledge of temporal technology, and the features of DBMS selected for the implementation [3]. These issues are primarily related to the problem of information security. It is difficult to ensure confidentiality, integrity and data accessibility, which semantics and features are difficult to describe by means of a RDBMS, where the Security System Mechanisms will operate.

3. Temporal Data Model

Currently in literature, 5 models of TDB are the most frequent [1,2,8]. The models differ by the type of relationship, and by the method of data presentation. Analysis of the models described in the literature, as well as some practical implementations can identify a number of types of temporal relations. The most common of them can be divided into the following three categories:

• By the type of ranges:

- Relations with disjoint time ranges;
- Relations with overlapping time ranges;
- Relationships that do not support versioning of objects.

• By a set of temporal attributes:

- Relationships with specific temporal attributes (days of the weeks, day parity, shifts, etc.);
- Relationship with a temporary attribute (start date, end date or the date of fixation). This may be a regular relational relationship that contains a temporary indicators, inherent to the object in life (invoice date, time of fixing the parameter values from sensors, the date of receipt, the date of termination, etc.);

• The ability to store incorrectly entered data;

- Supported.
- Is not supported.

Indicators of integrity, as well as a means of achieving them depend on the type of temporal relationship and the method of data presentation.

3.1. Mathematical Description of the Model

Let us first describe the temporal data model that is used to store objects of temporal system, to which is going to be applied methods of information security. As an optimal data model we suggest to use an open model with an Abstract Object Identifier (AOID), shown in Figure 3. In this model the life cycle (lifespan) of O is described through the life cycles of all its properties, defined in different ways and have temporary attributes $\boldsymbol{T}$ ={ $\boldsymbol{T_{start}}$, $\boldsymbol{T_{end}}$ } defining the start and end time of the life cycle.

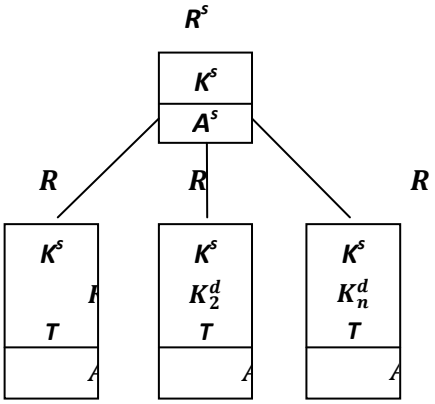


Fig. 1. Data model with an abstract object identifier

It is assumed that the temporal system's object $\boldsymbol{O}$ is characterized by the unique abstract identifier $\boldsymbol{K^s}$ and a set of attributes $\boldsymbol{A}$ and can be represented as:

$$\boldsymbol{O} = \{\boldsymbol{K^s}, \boldsymbol{A}\}.$$

Set of attributes $\boldsymbol{A}$, in turn, divided into a set of static attributes, $\boldsymbol{A^s}$, are not subject to changes over time, and a set of dynamic attributes $\boldsymbol{A^d}$, time-varying, ie

$$\mathrm{A} = \mathrm{A^s} \cup \mathrm{A^d}$$

The model will not consider multiple attributes, which are present in the sets $\boldsymbol{A^s}$ and $\boldsymbol{A^d}$, as it is considered by us in this context, and it does not matter, and will only lead to unnecessary complication of the model.

Since the values of attributes in relational systems must be atomic, then to represent the dynamic attributes of one relational relation is not enough. We represent the object $\boldsymbol{O}$ as a set of interlinked relational relations:

$$\mathrm{O} = \{\mathrm{R^s}, \mathrm{R_1^d}, \mathrm{R_2^d}, \dots, \mathrm{R_n^d}\}$$

Where $\mathbf{R_s}$ = $\{\boldsymbol{K^s}, \boldsymbol{A^s}\}$ - parent relationship, describing the absolute identity $\boldsymbol{K^s}$ and the object containing the static attributes of the object $\boldsymbol{A^s}$.

$\boldsymbol{R^s, R_1^d, R_2^d, \dots, R_n^d}$ - relations that describe the discretedynamic changing of attributes $\boldsymbol{A^d} = \{\boldsymbol{A_1^d}, \boldsymbol{A_2^d}, \dots, \boldsymbol{A_m^d}\}$ and having schema:

$$\boldsymbol{R_i^d} = \{\boldsymbol{K^s}, \boldsymbol{K_i^d}, \boldsymbol{T}, \boldsymbol{A_i^d}\}, \mathrm{i} = 1,2, \dots, \mathrm{n},$$

Where $\boldsymbol{T}$ = {$\boldsymbol{T_{start}}$, $\boldsymbol{T_{end}}$, $\boldsymbol{T_{tr}}$} - a vector of attributes of time, $(\boldsymbol{K^s}, \boldsymbol{K_i^d}, \boldsymbol{T_x})$Primarykey, where $\boldsymbol{T_x}$ - any attribute of $\boldsymbol{T}$, $\boldsymbol{A_i^d}$- discrete dynamic attribute, which is a part of $\boldsymbol{A^d}$, $\boldsymbol{K_i^d}$ - part of the primary key of $\mathrm{R_i^d}$, resolution non temporal relation "one to many" for Rs. In the case of ,$\boldsymbol{K_i^d} = \mathbf{0}$ non temporal link would be "one to one."

4. Principles of Information Security in The Temporal Systems

The purpose of information security is to provide authorized users access to relevant information and the certainty that this information is correct and that the system is available. These aspects lead us to the concepts of Confidentiality, Integrity and Availability (**CIA**).

CIA -is the standard abbreviation for the confidentiality, integrity and availability, are an essential components of information security (ISO 17799) (Fig. 2).

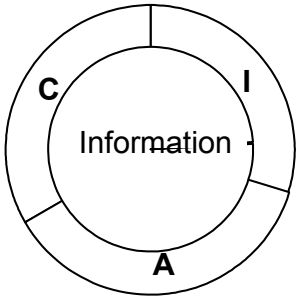


Fig. 2 The relationship between Information Security and the components of the model CIA

The concepts of confidentiality, integrity and availability are set out in the public literature [1,6,7], but when dealing with temporal systems the author found it necessary to extend the definitions in the context of temporal systems. In the context of temporal data security we will consider the concept of confidentiality, integrity and availability, stated below.

Confidentiality - protection in time of temporal data from unauthorized access, review and edit transactions, as well as the transfer of authority without the permission of the individual.

Accessibility - the ability to use the information in the TDB, when required. In other words, a state of the data when it is in the form, in the place and on the time required by the user. Access Tools to provide easy access to temporal information for authorized operations.

Integrity - protection of historical information from unauthorized changes, ensuring its consistency and completeness. Integral part includes a set of rules that maintain databases of temporal data in a consistent state, as well as protect it from losses in the reference.

The standard security control tools in RDBMS model is based on a discrete model, ie, user access rights management is possible on the level of named objects.For example, you can restrict user access to various operations to access the table or even its columns, as they both have their names. The problem is concurrent access to tuples, which, have no names. A tuple is characterized by life-cycle of an object or its properties, there for, it is precisely the quantity of information to which you want to restrict or allow access in TDB. Regarding the real-time, installation of confidentiality and availability of the tuple of temporal relationships are dynamic, while setting limits on access to RDBMS are static values.

To solve the above problem to control access to a relationship tuple an access restrictionmethod is used to restrict access relational tuples.

Let us consider the general problemsformalization of information security components in TDB, as well as a number of specific examples.

In describing the performances of these tasks we will use the temporal model described above in P. 3. In addition, we introduce the following notation characterizing the system of information security:

$\boldsymbol{U} = \{\boldsymbol{u}\}$ - a group of registered users in the system;

$\boldsymbol{G} = \{\boldsymbol{g}\}, \boldsymbol{g} = \{\boldsymbol{u}\}$ - a set of registered user groups;

$\boldsymbol{Op} = \{\boldsymbol{op}\}$ = {select, insert, update, delete} - A set of possible operations to access the tuple;

$\boldsymbol{F} = \{\boldsymbol{f}\}$ - set of functions of the security policy defined for the $\boldsymbol{Op}$.

Confidentiality problem - In general, the problem of Confidentiality can be formulated as follows: the user gets an access to the tuple, if the value of the function ofconfidentiality sign calculationof the tuple is true. This function contains one or more rules of the security policy defined for the object relations R.

The function for computing confidentiality sign of data looks as follows:

$$F_{conf}(R) = \begin{cases} \text{true, if } f\big(C(u \times op \times Now \times R)\big) = \text{true} \\ \text{false, otherwise} \end{cases}$$

where ***Now***- Current Time; ***op*** - Access Operation, in this case, the select operation uses "***select***"; ***f*** - logical function that combines ***C (u × op × Now × R)*** - a set of rules of confidentiality, taking into account the current time, user, and semantic rules of confidentiality, defined within ***R***.

The problem of integrity of temporal databases is to meet a specific set of integrity rules ***I***:

$$F_{integrity}(R) = \begin{cases} \text{true, if } f\big(I(u \times op \times Now \times R)\big) = \text{true} \\ \text{false, otherwise} \end{cases}$$

The problem of "garbage" in a temporal database lies at the junction of two problems: ensure the integrity and availability. The presence of "garbage" in the database makes it difficult to access relevant data, which affects their availability, and the "garbage" itself, by definition, is against the concept of integrity.

Under the "garbage" in the TDB will be understood tuples of temporal relationship that will never be used for reading or changing the system. "Garbage" can be defined segment of the object life "lifespan", which carries no special meaning and is completely contained in the beginning and end of the cycle in another segment with a later time lifespan transaction.

Let $\boldsymbol{R} = \{\boldsymbol{K^s}, \boldsymbol{A}, \boldsymbol{T}\}$ - temporal relation having attributes $\boldsymbol{A}$ and time attributes of $\boldsymbol{T}=\{\boldsymbol{T_{start}}, \boldsymbol{T_{end}}, \boldsymbol{T_{tr}}\}$ and containing n tuples:

R = R1 □ R2 □ ... □ Rn.

Then the tuple $\boldsymbol{R}$ □ $\boldsymbol{R}$ is a "garbage", if in the temporal relation $\boldsymbol{R}$ there is another tuple of $\boldsymbol{Ri}$□$\boldsymbol{R}$, defining the standard segment of the same object as an $\boldsymbol{R_1}$, but with the later time of the transaction, and the life period of a tuple of $\boldsymbol{R_1}$ is completely covered vital segment of the tuple $\boldsymbol{R_i}$:

$$F_{Garbage}(R_1) = \begin{cases} \text{true,} & \text{if } \exists R_i \in R \vee R_1(K^s) = R_i(K^s) \vee \\ & [T_{1start} : T_{1end}] \subset [T_{istart} : T_{iend}] \vee \\ & T_{itr} > T_{1tr} \\ \text{false,} & \text{otherwise} \end{cases}$$

Problems of integrity in TDB include also problems of redundancy, wordiness, and data inconsistencies. Their description and solution are addressed in a number of papers [1,8]. But the presented results in general are applicable to the relations with disjoint time frames, whereas for a number of practical problems is not always possible to be limited with such type of relationship. There are problems that require to operate on intersecting life cycles of objects [3,4]. Issues of integrity for this relationship need special consideration.

5. Implementation Levels of Information Security in TDB.

If we look at the geography of a simplified information flow from the storage system to the user, then the implementation of information security can be placed at different levels of the chain (see Fig. 3)

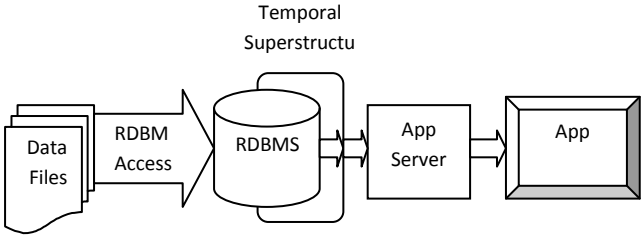


Fig. 3. The chain of information flow in the temporal system

Unfortunately, experience shows that often, the level of security implementation is the level of the client application. In this case, the security system has obvious

flaws, because the user can access the data, bypassing the application, and another program that works with the same set of data can have a different idea of the information security of temporal data or do not have an idea at all.

The level of server applications can also be a good solution, because at this level with the object languages can be very simple and flexible to describe all the nuances of the security system. At the level of DBMS, or corresponding temporal structure, the capability of flexible descriptions can be much lower because we have to be limited to lingual features of DBMS and a relational presentation of data. The disadvantage here - is the need to restrict access to data through another server application. To provide an exclusive access to data using only one server application, that is not always possible in a commercial operation. And to implement and continuously maintain a common understanding of security system in two different applications is very difficult and hopeless task. The most preferred level, according to the author, is the level of temporal superstructure. In this case, it has the following advantages: DBMS already "knows" the semantics of temporal data and the security system uses this knowledge. Neither user program nor data access via SQL commands will be able to bypass this protection.

6 Conclusions

Information Security for temporal data is based on the well-known triad: confidentiality, integrity and availability. Provision of these indicators in temporal environments endowed with many special features. An example of the need for confidentiality, and availability may be limiting access to data for users for some time periods and allow access to the others; integrity - ensuring the temporal normal form, or the condition of a time gap in the system, for example, in the event of some change in the on the schedule of passenger train schedule system. Providing similar conditions with temporal data is not part of the standard mechanism for information security in a relational DBMS. Moreover, in the specialized temporal DBMS that has received little attention.

The work is complicated by the fact that in temporal systems there are several types of temporal relations, and methods for providing information security of each type may be different. The paper contains a formalization of some problems of information security for the specified types of temporal relations, constructed with reference to problems of rail transport. Direction for further research are the analysis of the characteristics of temporal relations with overlapping time periods and the development of methods to provide the confidentiality, integrity and availability of such relations.

7 REFERENCES

[1] Date C.J., Hugh Darwen, Nikos A. Lorentzos. Temporal Data and the Relational Model: A Detailed Investigation into the Application of Interval and, Relation Theory to the Problem of Temporal Database Management. Morgan Kaufmann, December 2002, 480 p.

[2] B. Galatenko Information Security Standards: The course of lectures. Moscow: INTUIT.ru (Internet TrackInside Inc Information Technology), 2004. 328 p.

[3] Kopitovs J., Demidovs V., Petoukhova N. Method of Temporal Databases design Using Relational Environment. In: Scientific proceedings of Riga Technical University - Computer Science: Applied Computer Systems, Series # 5, Issue # 13. Riga: RTU, 2002. pp. 236-246.

[4] Petoukhova N. Development of a Complex Security System in Relational Databases for Railway Transport. In: Scientific proceedings of Sixth International Baltic Conference on Computer Science and Information Technologies - Databases and Information Systems: DB & IS 004. Scientific Papers University of Latvia. June 6-9, 2004. Volume 673. pp. 139-150.

[5] Preliminary information for International Symposium on Secure Software Engineering ISSSE 06 - IEEE, http://www.ieee-security.org/Calendar/cfps/cfp-ISSSE06.html (2005, November 22)

[6] Microsoft Corporation with Andy Ruth and Kurt Hudson. Security + Certification Training Kit, Microsoft Corporation, 01/29/2003, 512 pages (CIA triad, pp. 5-6, 11)

[7] Vyukov N., V. Galatenko Security database management systems. Systems database management, # 1/1996. 1996, Open Systems Publishing House, Moscow. S. 29-54

[8] Jensen C. Temporal Database Management. Dr. techn. Thesis, defended on 14.04.2000, 1328 pages, http://www.cs.auc.dk/ ~ csj / Thesis / (2005, November 22)

[9] N. Petukhov method of providing access to relational data at the row level relationships.Transport and Telecommunication. Vol. 4 (1). 2003, Riga, Latvia. S. 45-52.

QXC-An efficient Querable XML data Compression method in EMR system

Xiaoyuan Bao [1], Wen Zhao [2], Yanzhong Jin [3], and Shikun Zhang[2]
[1] Computer Science Department, Peking University, Beijing, China
[2] Software Engineering Center, Peking University, Beijing, China
[3] Computer Science Department, Tianjin Technology & Science University, Tianjin, China

Abstract *- The use electronic medical record (EMR) systems in medical facilities become a standard practice in the medical care industry. In the system, each patient's record represented as XML trees intend for easy information exchange among doctors and related researchers. But as hundreds of thousands patients visits and events information is stored in the system as XML format, query some important info such as diagnosis detail (especially keyword query with XPath) becomes very slowly, partly for massive XML storage, and partly for the complex XML trees structures. How to store and query XML data efficiently in our system PKUEMR is the motivation of our work.*

In fact, as for our best knowledge, existing XML compression methods proposed are not efficient for storage and query/retrieval evaluation of XML data. We proposed an XML structure reserving compression method QXC, using QXC we can compress and decompress XML data efficiently. In QXC, we compress XML structure and content using marked byte-oriented Huffman. The evaluation of XPath query with keyword leaf predicates(such as //p1//p2//…//pn[text]) in EMR system can be done directly on compressed XML data produced by QXC.

Keywords: XML,Compression, Huffman, XML Query

1 Introduction

The use electronic medical record (EMR) systems in medical facilities become a standard practice in the medical care industry. In the system, each patient's record represented as XML[1] trees intend for easy information exchange among doctors and related researchers. But as hundreds of thousands patients visits and events information is stored in the system as XML format, query some important info such as diagnosis detail becomes very slowly, partly for massive XML storage, and partly for the complex XML trees structures[3~7]. How to store and query EMRs efficiently in our system PKUEMR (Peking University Electronic Medical Record) is the motivation of our work.

In fact, as for methods of XML compression proposed in [8~12], they are not efficient for storage and query/retrieval evaluation of XML data. Some existing methods can compress XML data effectively and that we must uncompress the whole compression file to query the data. As for our applications, if we adopt this kind of compression algorithms, things will become worse: we have to uncompress the data and then carry out the query evaluation, apparently need more time to fulfill our task.

At the same time, the mostly real requirements of our doctors and researchers are that the EMR system can find keywords text with semantic path Q=//cancer//admissions //diagnosis[contains("blood")](XPath[2]) very efficiently. If we can compress the XML data and evaluate the queries directly on the compressed XML data, then we will have better system performance.

Based on this consideration, we propose an XML structure reserving compression method, QXC (Querable XML Compression), with the XML structure reserved, we can evaluate the Path of the query above directly on the compression XML data. At the same time, for the text string in the XML data, we adopt byte-oriented Huffman methods to compress the text data for the purpose of compression and query key words data directly on the compressed text data. By combing the structure reserving XML tree compression and Huffman dictionary text data compression methods, the EMR system can process the queries as above efficiently.

2 Preliminaries

In fact, with the development of XML application in practice, how to store and query XML data in EMR system is a hot research focus nowadays. In XML, every semantic unit is tagged by the respective <Tag>…</Tag>, so the redundancy because of this makes it more necessary for us to compress XML data. In EMR, the most important thing we care about is the storage and query/retrieval efficiency, and there is no need to modify the data in EMR. High storage efficiency means that less storage space used, which can be implemented by compression. As for efficient XML data query/retrieval, if the operation can be applied on the compressed XML data directly, then the query/retrieval efficiency can be improved at a certain extent over the original un-compressed data. At the same time, the found content in compressed XML data must be decompressed to be retrieved by the end user and this requires that the decompression efficiency must be high enough.

There are two kinds of XML compression: structure-reserving and non-structure-reserving compression.

With the non-structure-reserving methods, the resulting data does not have the XML format anymore, querying on the compressed XML data directly is impossible. The decompression must be performed before the query evaluation can be done. A naive non-structure-reserving compression method is to compress XML data using familiar compression algorithms such as Lempel-Ziv[8]. Recently, XMill[9] and XComp[10] are representative methods of this kind of compression.

The compressed XML data has the same structure as its original content after compression by non-structure-reserving method. Query evaluation may be performed directly on the resulting XML data. XGrind[11], XPress[12] are representative ones of this kind. We have found the following problems of existing XML compression methods in our research and practice work: 1. Improving compression ratio by semantic content classification does not work so well in practice (XMill, XComp). In fact, for the same semantic content, people can use many forms to present it, how to exactly know that they have the same original meaning is an open problem until now. 2. The existing methods do not support querying on compressed data directly and wholly, and the decompression efficiency, which plays important role in EMR systems, is not so well. 3. The existing compression methods were designed and work independently by themselves and their implementation is non-trivial, integrating them in EMR system to support compressed XML data based query/retrieval is still an open problem, etc.

With above problems in mind, we propose an XML compression method: QXC (Querable XML Compression), this is a structure-reserving XML compression method which enables querying on compressed XML data directly and wholly, and the decompression efficiency is also better. The contributions of our paper are described as follows:

We propose QXC, an XML structure-reserving compression method, by which we can compress and decompress XML data efficiently.

We propose X-ByteHuff, the core compression coding algorithm, which is based on general Huffman coding. The implementation of this algorithm is very simple.

We analysis how to query on compressed XML data directly resulting from QXC.

We have performed lots of experiments and the results demonstrate that our method is efficient.

The rest of this paper is organized as follows: Section 2 describes details of QXC; Section 3 discusses the query evaluation directly over compressed XML data; Section 4 and 5 give experimental results and conclusion respectively.

3 Querable XML data Compression

3.1 Overview

In QXC, we compress XML structure and XML content in the same way. For input XML text stream, we leave structure taggers ("<>", "</>", "=" or so) as their original form. By this, the structure of XML can be reserved. For element tags, attribute names, element contents and attribute values, compress the input stream by assigning Marked Byte-Huffman code to each word in them. Following we describe the Marked Byte-Huffman code tree firstly (2.2), then we discuss how to compress XML efficiently with the aid of P-Tries tree (2.3).

3.2 Marked Byte-Oriented Huffman Code

The original method proposed by Huffman [13] is mostly used as a binary code, i.e., each symbol of the input stream is coded as a sequence of bits. In our method, the Huffman codeword assigned to each text word is a sequence of whole bytes. For finding the start position of each word in compressed XML data, we leave the MSB of each byte as mark bit, that is to say, for a sequence of bytes assigning to a word, the MSB of the first byte is set to 1 and others are set to 0. The Huffman degree of this method is 128, instead of 2.

In this paper, we consider that the words and separators of the text are symbols, and the separators are codified using the spaceless word model.

According to Huffman [13], We have the following equation for Byte-Huffman:

$$\frac{N-n_0}{D-1}=K$$

In which n0 is that the words counting number on the deepest level in Huffman tree. D is the degree of nodes in Huffman tree. Here its value is 128. Now we have

$$n0=N-K*127.$$

Now we perform mod operation on both sides by 127, then

$$n0= (N \bmod 127).$$

Now we propose an efficient way to construct the Marked Byte-Huffman tree. A table T_{bh} (T means both Table and Tree) is used to store the Marked Byte-Huffman tree. Each record in Tbh stores information about a word in the compression vocabulary, it composes of the following five fields:

ResultCode | Word | Frequency | Children | LocoalCode

ResultCode: the Marked Byte-Huffman code of the word;

Word: the word text string;

Frequency: the word count number after the first scan on XML text file;

Children: for an inner node in the Marked Byte-Huffman tree (T_{bh}), it is the pointer to its child which has the biggest frequency value;

LocalCode: the local order of the node in the sub-tree rooted from its parent node, where the node with bigger frequency value has a larger LocalCode number.

In T_{bh}, each level nodes of the Marked Byte-Huffman tree are stored continuously. The inner nodes of the tree are pointers to next level node, so the word field of inner nodes is empty. Figure 1 is the construction sample of T_{bh}.

Following in algorithm 1 is the construction of Marked Byte-Huffman tree.

Algorithm 1. Construction of Marked Byte-Huffman tree
Input: T_{bh}[1..(N+m), 1..5] //m=K+1, for K see step no. 3
Output: T_{bh}

1. Store N words of compression vocabulary with ascendant order of frequency in record 1, 2, …, N of T_{bh};
2. Field Word and LocalCode are set to 0;
3. Calculate n_0=(N mod 127); K=N/127;
4. LocalCode fields of records 1～n_0 are set to 0, 1, …, n_0-1, respectively;
5. Sum the value of frequency fields of first n_0 records, and put the result to record N+1, the Children value of this record is set to 1;
6. For i=1 to K do (6.1～6.5)

6.1 start_record=1+n_0+(i-1)*128; end_record=n_0+i*128;
6.2 Sort the records in T_{bh} from start_record to N+i on frequency value in ascendant order;
6.3 LocalCode fields of records from start_record to end_record are assigned number 0,1,…,127, respectively;
6.4 Sum the frequency fields value of records start_record～end_record, the result is stored in frequency field of record N+i+1;
6.5 Children field of record N+i+1 is set to 1+n_0+(i-1)*128;

7. End.

3.3 X-ByteHuff

We propose a vocabulary index structure P-Tries to accelerate the Marked Byte-Huffman codes searching during compression. P-Tries combines Patricia Tree [14] and Tries [15] together. In fact, P-Tries is a directed rooted tree. In P-Tries, there is a virtual root node. From root node, each edge is labeled with a distinct character which comes from the first characters of all words in the vocabulary. Each inner node has an attribute *pre* which is integer type, to denote the common prefix characters counting, of words in vocabulary, that is to say, there are words in vocabulary have common prefix of *pre* characters. Figure 2 is an example P-Tries tree.

Take words "angle" and "anger" as example. They have the common prefix "ang", so the branch labeled "a" (the first character of "ang") directed to a node which *pre* attribute is 3. From there on, "angle" node branches with edge *l* and "anger" with edge *e*, the corresponding leaf node stores the word string and its length. There is another attribute we add to each leaf node: *HuffPointer*, which points to the corresponding record in T_{bh}. By this, when we find a word in P-Tries, we get the corresponding Marked Byte-Huffman code at the same time. Finding a word *w* in P-Tries is described as follows:

1. Travel edge *w*[1] (the first character) from the root node (such an edge must exists), and we arrive at node *y*;
2. If node *y* is an inner node then travel from it along the edge *w*[*y.pre*], and go to step 3; else go to step 4;
3. Continue step 2 in the same way continuously;
4. If *w*=*y*.word then retrieve its compression code *y.HuffPointer*->ResultCode.

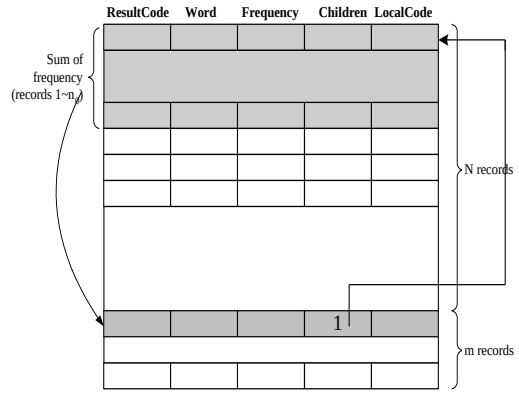


Figure 1. Construction of T_{bh}

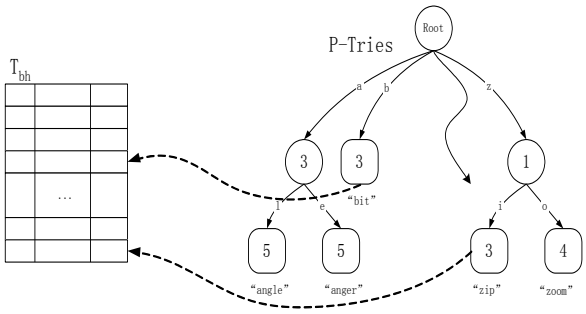


Figure 2. An example P-Tries tree

Now it is time to discuss the XML compression using Marked Byte-Huffman coding, named X-ByteHuff.

XML structure compression

We leave structure taggers, including "<", ">", and "/" etc., as its original form in the compressed text. By this, the XML structure is remained, and some existing XML processing tools such as SAX can be applied directly. As for element tag strings, attribute name strings, decompose the strings into words and put these words in T_{bh} for code construction. In the compression, start tag of each element will be compressed to have the form of <$TAG_{\text{Byte-Huffman}}$>, where $TAG_{\text{Byte-Huffman}}$ is the Marked Byte-Huffman code of the corresponding tag string (ex. <0x12> for <Author>), for end tag, copy "</>" to the compressed stream. We also reserve "=" for separating compressed attribute name and its value.

XML content compression

For the text content of elements and value of attributes, replace each word with its corresponding Marked Byte-Huffman code for writing out to compressed stream. We have mentioned that we use MSB of the first byte in compression code as the mark of a word start, we implement this by perform a bit-oriented *or* operation on the first byte with 0x80 during the construction of T_{bh}.

In all, we give the X-ByteHuff algorithm for compressing XML data with structure reserved.

Algorithm 2. X-ByteHuff
1. Scan XML data for the first time to put words and their counting number (frequency) in T_{bh};
2. Produce ResultCode of each word; (recall that must perform *or* operation on the first byte)
3. Construct vocabulary index structure P-Tries;
4. Read each word in input stream, find the respective Marked Byte-Huffman code in T_{bh}, with the aid of P-Tries, and put the code (byte sequence) in the output stream. In the process, leave structure taggers untouched and copy them directly to output stream.

4 Query evaluation analysis

In this section, we will discuss how to evaluate query on compressed XML data directly.

Firstly, we explain the reason why we can query directly on compressed XML data. According to X-ByteHuff, the compression is implemented as two main steps: the first step is to count the words in source XML text file and generate respective Marked Byte-Huffman code of each word; the second step is to replace each word with its compression code. That is to say, a word (ex. *w*) is compressed with the same code, independent of its position in the source XML file. By this, when we have to find *w* in compressed data, we can searching in P-Tries to get its compression code w_c firstly, and because we have already marked the start of each compressed word, so we can scan and match w_c in compressed data word by word to get all the results.

Above is a simple explanation for finding a word in compressed data. For XML data in IR, we always use XPath queries of form (*Q*=)//element-$_1$/element$_2$/…/element$_m$[text()=*string*]. The query is to find all the *element*$_m$ which has the text content *string*, and its parent element is *element*$_{m-1}$, which in turn has the parent element *element*$_{m-2}$, and so on. We call //element$_1$/element$_2$/…/ element$_m$ of *Q* the *structural query part* and [text()=string] the *content query part*. Following we discuss how to evaluate *Q* on compressed data directly.

For simplicity reason, we suppose each element tag contains only one word. Query *Q* must be re-written as //H_{e1}/H_{e2}/…/H_{em}[text()=H_{string}], where H_{ei} is the compression code of element$_i$ tag string (word), i=1,2,…,m, and H_{string} is the compression codes stream of *string*.

Because we compress XML data with structure reserved, the structural query part evaluation (//H_{e1}/H_{e2}/…/H_{em}) can be processed based on its XML data tree of compressed text in a naive way: traveling XML data tree in depth first order (or width first order) to find an element H_{e1} firstly, from H_{e1}, finding if a child element H_{e2} exists, which in turn find if a child element H_{e3} exists and so on, until we arrive at H_{em}, then we judge if its content is H_{string}. If all these steps successfully finish, then we have found a matching one (H_{em}) in the compressed XML data. For all H_{e1} in XML data tree, do the above process and finally we get all matching element(s). Please note that we only give a simple evaluation method here to prove that we can query directly on compressed XML data, for *structural query part* evaluation in practice, people always use some index structure on compressed data to improve the query efficiency.

In EMR, most cases of *content query part* has the form of [text()$*string*], which means that the content of an element contains *string*, and as a convention, the words in *string* can be matched un-continuously. With this in mind, we use an NFA based method [16] to process the *content query part* of *Q*. Figure 3 is used to explain how to fulfill the matching.

Take string "road rose is" as example. There are three words in *string*, so the driving conditions in NFA are 3 bits long. And we add a mask field in T_{bh} to store each driving condition bits string of the corresponding word in *string*.

Firstly, we searching "road" in P-Tries to arrive at record of "road" in T_{bh}, because it is the first word in string, so the mask field of this record is set to "1xx"; next, we search "rose" to set the "rose" record to "x1x"; and finally set the "is" record to "xx1", "x" means it can be 0 or 1.

Now suppose that we have already arrived at a H_{em} element in XML data tree, then we read the compressed content of element H_{em} byte by byte, during the process we travel the Marked Byte-Huffman tree (T_{bh}) from root node to leaf node, along the edges labeled with these bytes. For example, if we read the compression code of "road", we can travel the T_{bh} to the corresponding leaf node which has the mask field value "1xx", and using this mask value to drive the NFA from state s0, because it is an effective driving condition, then state s1 is activated, means that we have matched the first word of *string*. The following steps perform in the same way and finally when we get to the final state s3 of the NFA, it means that we have matched *string* (actually is H_{string}) in the compressed content of H_{em}.

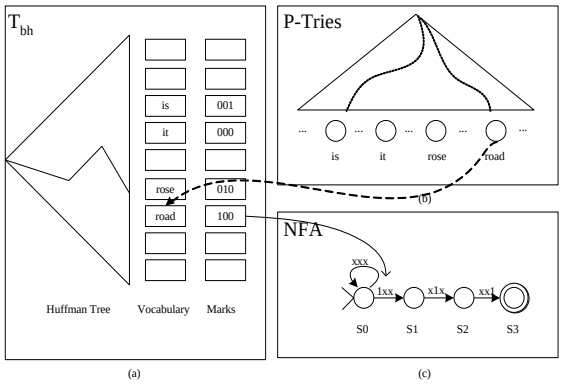
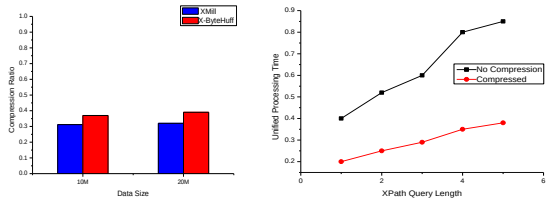


Figure 3. Matching H_{string} in the content of H_{em}

5 Experimental results and analysis

Figure 4 is the experimental results. In (a), we compare the compression ratio of QXC with XMill, we can see that XMill outperforms QXC by 4% ~ 6%, for the reason of supporting query directly, we consider that this weakness is within our tolerance. In (b), we compare the querying time on

compressed data with original text. We can see that querying on compressed data directly outperforms querying on original text at an average of 2 or above times more quickly. In the experiments, we use DBLP [17] as our experimental data set, under Windows XP, Intel dual core CPU 2.8GHz, 512MBytes memory and 120GBytes hard disk.



(a) (b)

Figure 4. Experimental results

6 Conclusion and future works

We proposed an XML compression method QXC, querying can be evaluated upon the compressed data directly. We will focus our future work on complex query evaluation methods.

7 Acknowledgment

This research work is supported by National Core-High-Base Major Project of China under Grant No.2010ZX01045-001-008 and National Grand Fundamental Research 973 Program of China under Grant No.2009CB320706. Thanks for all the people helping me in this research.

8 References

[1] Dan Suciu: Semistructured Data and XML. FODO 1998:1-12

[2] XPath: http://www.w3.org/standards/techs/xpath

[3] Robert C. Hsiung: Adoption of electronic health records by medical specialty societies. JAMIA 19(1):143 (2012)

[4] Michael Farnum, Victor S. Lobanov, Frank Defalco, Soledad Cepeda: Opening the Door to Electronic Medical Records: Using Informatics to Overcome Terabytes. BIBM 2011:659

[5] Vishesh Ved, Vivek Tyagi, Ankur Agarwal, Abhijit S. Pandya: Personal Health Record System and Integration Techniques with Various Electronic Medical Record Systems. HASE 2011:91-94

[6] Wei Xu, Zhiyu Guan, Hongxin Cao, Haiyan Zhang, Min Lu, Tiejun Li: Analysis and evaluation of the Electronic Health Record standard in China: A comparison with the American national standard ASTM E 1384. I. J. Medical Informatics (IJMI) 80(8):555-561 (2011)

[7] Jun Liang, Mei Fang Xu, Lan Juan Li, Sheng Li Yang, Bao Luo Li, De Ren Cheng, Ou Jin, Li Zhong Zhang, Long Wei Yang, Jun Xiang Sun: Increasing the Meaningful Use of Electronic Medical Records: A Localized Health Level 7 Clinical Document Architecture System. ADMA 2010:491-499

[8] J.Ziv, A. Lempel. A universal algorithm for sequential data compression, IEEE Transactions on Information theory, 23(3), 1977, 337~343

[9] H. Liefke, Dan Suciu. XMill: An efficient Compressor for XML data, Proc of the ACM SIGMOD Int'l Conf on Management of Data, 2000, 153~164

[10] Weimin Li. XCOMP: AN XML COMPRESSION TOOL, Master Thesis of Waterloo University, Ontario, 2003

[11] P. M. Tolani, et al. XGRIND: A Query-friendly XML Compressor. Proc of the 18th Int'l Conf on Data Engineering (ICDE' 02), 2002, 225~225

[12] J.-K. Min, et al. XPRESS: A Queriable Compression for XML data, Proc of the ACM SIGMOD Int'l Conf on Management of Data, 2003, 122~133.

[13] David A. Huffman. A method for the construction of Minimum-Redundancy Codes, Proc. Of the Institute of Electrical and Radio Engineers (I.R.E), 40(9), 1952, 1090~1101

[14] Donald R. Morrison. PATRICIA-Practical Algorithm To Retrieve Information Coded in Alphanumeric, Journal of the ACM, 15(4), 1968, 514~534

[15] Donald E.Knuth. The art of computer programming, vol.3, Tsinghua University Press, 2002

[16] E.S. de Moura, Gonzalo Navarro, et al. Fast Searching on Compressed Text Allowing Errors, SIGIR'98, 1998.

[17] DBLP, XML Data, http://dblp.uni-trier.de/xml

Self Organizing Maps: A Robust Implementation

John Wittenauer and Ray Kresman

Department of Computer Science
Bowling Green State University
Bowling Green, OH 43403

Abstract

Methods for visualizing multidimensional data are of great interest in computer science and engineering. One popular technique is self-organizing map, a type of neural network, that uses machine learning algorithms to map multidimensional data to a two-dimensional surface. They are widely used for exploratory data analysis and visualization and have been used to perform clustering and classification tasks successfully. This paper builds a robust and extensible self-organizing map implementation capable of producing several visualizations and evaluates the quality of the maps that it generates.

1 Introduction

Self-organizing maps (SOM) have been used successfully all over the world for a myriad of different applications. Since the SOM can be useful for any sort of data analysis, the possibilities are nearly endless. A few such notable applications of self-organizing maps have included domains such as speech recognition, control engineering, biomedical sciences, and financial analysis [4]. The self-organizing map is a single-layer feed-forward neural network that uses an unsupervised competitive learning algorithm to build a topology-preserving model of the input data [3], [4]. The goal of the self-organizing map is to reduce the dimensionality of a data set to discover some underlying structure. The algorithm takes a set of n-dimensional input vectors and produces a two-dimensional discretized representation of the input space. The nodes of the map (also called output nodes) are fully connected to the input; in other words, each input vector is connected to each output node. The map is said to be "topologically-preserving" because vectors that are "similar" will end up close to each other on the map. This occurs because during the competitive learning stage (in which a "winning" node that is closest to the input vector is selected), the algorithm not only updates the winning node's weight but also updates the weight of neighboring nodes. This has the effect of "pulling" neighboring nodes towards the winning node and, after a suitable number of iterations, results in the map's topology. The original self-organizing map algorithm was developed by Teuvo Kohonen [1, 2] and is still the most popular incarnation of the technique.

Given the popularity of self-organizing maps, there have been a number of freely available implementations. The most notable implementation is SOM_PAK [6], developed in part by the original creator of the self-organizing map, Teuvo Kohonen. Although SOM_PAK is both lightweight and efficient, it does have some shortcomings for a modern application. First, it uses a command-driven interface rather than a graphical interface. The number of available commands is quite extensive, and while it is a powerful method of interaction for an expert user, it comes with a steep learning curve. Another issue is that the package does not include any built-in visualization tools. It instead allows a user to save a trained map and employ visualization techniques independently using the map's weight vectors. This design is conducive to flexibility and platform independence, but is not a good way to learn how the map training process works or how the various parameters affect the final outcome. Finally, SOM_PAK was written in the C language for the purpose of performance and efficiency. While it certainly achieved that goal, the choice of language somewhat limits the readability of the implementation.

The goal of this paper was to build a self-organizing map implementation based on Kohonen [1] that remains lightweight and efficient but additionally solves the three issues mentioned above by providing a robust graphical user interface, built-in visualization tools, and a well-documented code base using a modern object-oriented language. By contrast, the application developed in this paper is restricted in the sense that it requires the .NET framework to run. The application is highly configurable through an intuitive graphical interface so that many of the parameters of the algorithm can be controlled graphically without a steep command-

based learning curve. A visualization panel is also built in with several visualization techniques already implemented to give the user feedback on a trained map. This visualization can even be updated in real time as the training occurs to show the user how the map evolves. Additionally, it provides statistical information about the map's quality and training time.

2 Kohonen's Algorithm

The algorithm by Kohonen [1] can be summarized at a high level as follows. First, the weights for each output node must be initialized. Once this is done, the training begins. For each training iteration (also called an epoch), an input vector is chosen at random from the set of input data. The input vector is compared to the weight of each output node to find the winning node, also known as the best matching unit (BMU). Once the BMU has been found, all other units nearby are found via the neighborhood function. The winning node's weight is adjusted to be more like the input vector. Similarly, all of the nodes in its neighborhood also have their weights adjusted (the degree of adjustment depends on how close the node is to the winning node). This process is repeated for however many epochs are necessary until the map is complete. Now let's look at each part of the algorithm in more detail.

For the initialization phase, the weight vector of each output node must be set to some initial value before training begins. Each output node has two components associated with it – a topological location (for example (0, 0) or (3, 5)) which does not change throughout the training process, and a weight component for each input attribute or dimension. The topological location of a node is set when the map is constructed, so initialization refers to setting the weights of the node that are adjusted during training. The initial values do not really matter; they can be set randomly, the only important factor is that they are not all the same. Sometimes the weights are normalized, but this is not strictly necessary. Once initialization is done the main loop of the algorithm can begin.

For each iteration of the main loop, an input vector is chosen at random from the set of input data. The input vector is compared to the weight of each output node to determine a winner, or BMU. The BMU is defined by the output node that is "closest" to the input vector. Although there are a number of methods that have been used to determine this, the most popular is a straightforward Euclidean distance metric. For each input vector V and output node weight vector W, the Euclidean distance between them is defined as

$$Dist(V,W) = \sqrt{\sum_{i=0}^{i=n}(V_i - W_i)^2} \qquad (1)$$

where V_i and W_i are the components of the input and weight vectors, respectively.

Once the BMU is found, the next step is to calculate which other nodes are in the BMU's neighborhood. This is based on a neighborhood function that defines a radius around the winning node (based on topological location in the lattice). The next step is to update the weights of each node in the neighborhood (including the BMU itself). The weight vector adjustment can be defined as

$$W(t+1) = W(t) + \sigma(t)\alpha(t)(V(t) - W(t)) \qquad (2)$$

where t is the current time-step, $W(t)$ is the current weight vector, $V(t)$is the current input vector, $\sigma(t)$ is a distance function that reduces the influence of the weight adjustment at greater distances from the BMU, and $\alpha(t)$ is the adaptation gain or "learning rate" of the weight adjustment. The choice of $\alpha(t)$ and $\sigma(t)$ can vary; as with the neighborhood function, they may be linearly or exponentially decreasing.

After the weight adjustments have been carried out, the whole process repeats. The number of time-steps necessary will vary depending on the number of input samples, but is generally in the thousands. A rule of thumb proposed in [1] is that the number of iterations should be at least 500 times the number of output nodes. As noted above, there are several functions in the algorithm that may vary by implementation (neighborhood function, learning rate, and distance decay function). In addition, there are a number of parameters involved in building a map such as the number and arrangement of output nodes or method of initialization. See [1] for complete details.

3 Visualization Methods

Once a map has been trained on a data set there are a number of ways that it can be used. Extensions to the original self-organizing map algorithm are able to detect clusters of similarity in the map, and if the training data has a known classifier associated with it then the map can be enhanced to classify new untrained instances of the data just as any other classifier would be able to do. However, the most common use of a trained map is as a visualization tool. There has been a lot of research into techniques for visualizing a trained SOM using color and other graphical means, and as a result there are quite a few techniques that have been

developed with varying (and sometimes situational) degrees of usefulness..

3.1 Blended View

One way to visualize the map is to assign a color to each dimension in the map vectors. The simplest example would be using three-dimensional input data and assigning red, blue, and green to each dimension, respectively. However, this does not work for higher-dimensional data so a clever approach is needed. Schatzmann [4] demonstrated a method in which the HSB (hue, saturation, brightness) color wheel is divided into *n* evenly-spaced partitions where *n* is the number of dimensions in the map vectors. This gives us a color C_i for each dimension such that

$$C_i = C_i[Red], C_i[Green], C_i[Blue \quad (3)$$

We can now obtain a color C_v for each map node with vector $v = (v_1, v_2, \dots, v_n)$ where

$$C_v[Red] = \frac{\sum_i C_i[Red] * v_i}{\sum_i C_i[Red]} \quad (4)$$

with identical equations for the green and blue components. The result is a composite color at each node in the map. This creates a "blended" view of the map with similar nodes having relatively similar colors. This technique is useful for observing very general trends in the data but it is usually not possible to derive much specific knowledge from this view, particularly when the number of dimensions is very high.

3.2 U-Matrix

Another visualization method is called the unified distance matrix, or U-matrix. The U-matrix is constructed on top of a trained map and calculates the average distance (in vector space, not topological distance) from a node to each of its neighboring nodes [5]. More formally, if we define *n* as a node in the map and *adj(n)* as the set of neighboring nodes to *n* then

$$U(n) = \sum_{m \in adj(n)} Dist(n, m) \quad (5)$$

where *U(n)* is the U-matrix value for node *n*. Once this has been calculated for each node in the map it can be displayed visually, usually as a grayscale image. The end result is a graphic that shows the landscape of the data in terms of similarity to surrounding nodes. Clusters of white show areas where the vectors are very similar to each other, and dark bands show gaps in the landscape where vectors are changing very quickly. Such a technique is often used in conjunction with the blended view to get a feel for the overall landscape of the data. This method is also commonly used in conjunction with a clustering algorithm if clustering is to be performed on the data.

3.3 Component Planes

Both of the above techniques attempt to visualize the entirety of the map in one image. Often this means that information is going to be lost, especially if the data is very complex. Another approach to visualizing a trained self-organizing map is to look at each dimension in the map separately. This view, called a component plane view, creates a separate image for each dimension in the map vectors. The color of the map scales with the value of the dimension being visualized at the given node. An important feature of the component plane view is that the maps are still topology-preserving, meaning that correlations can be drawn between different component planes by using the same topological location in the map. However one caveat of this technique is that it gets very cumbersome as the number of dimensions in the data increases. It also typically requires some knowledge about what the dimensions represent in order to derive useful information.

3.4 Quantitative Evaluation

One issue with visualization techniques is that their results are qualitative. We cannot measure the "correctness" of a colored image as a representation of the original input data. For this reason it is necessary to quantitatively evaluate a trained map as well to confirm that the map is representing actual characteristics of the input data. One such measurement proposed by Kohonen [2] is the average quantization error (AQE) of the map. AQE is a measurement of how closely the nodes in the map represent each instance of the input data. The quantization error for any particular instance is calculated as the distance (see Equation 1, Section 2) between that instance and the best-matching unit on the map. The final AQE is a simple average of the quantization error for each instance in the data. Stated more formally, we define the AQE as

$$AQE = \frac{1}{n} \sum_{m \in data} Dist(m, BMU(m)) \quad (6)$$

where *n* is the number of instances in the data and *m* is an instance in the set of input data. See [1] (also Section 4.1 below) for details on BMU.

4 Design and Implementation

The platform chosen for this project is the .NET framework 3.5. The code is in C# 3.0 and it was built using Visual Studio 2008. The implementation is constructed using three classes for the algorithm plus a fourth to manage the user interface. The three classes associated with the self-organizing map algorithm are the parser, cell, and map classes. The parser class is responsible for reading in an input file and converting the input data into a two-dimensional array of normalized floating-point numbers. Each column in the array corresponds to an attribute or dimension in the input data, and each row is a separate instance of the data. The cell class provides the structure for a node in the map and has properties for its map position and its associated weight vector. The cell class also provides an overloaded function to calculate its vector space Euclidean distance from either an input vector or another cell in the map. The map class brings everything together to create a map that can be trained using the self-organizing map algorithm. Each map instantiation has one parser and a two-dimensional array of cells as properties. The map class also has numerous properties to manage various configurable parameters in map training as well as a two-dimensional-array of floating-point numbers to store the U-matrix of a trained map. Methods provided by the map class include an average quantization error calculation, a U-matrix calculation, and a map training method that executes one iteration of the SOM algorithm. Also included is a private function that finds the best matching unit of an input vector to the current map. The form class contains code for the interface and button events as well as methods for each of the three built-in visualization panels. This class also contains private methods to calculate red, green and blue scaling color values for the component planes visualization. For brevity other details of the implementation are omitted in this discussion.

4.1 Map Class

The bulk of the application code related to the functioning of the algorithm is contained in the map class. The map class also contains three private functions. The first one, CalculateAQE(), calculates the average quantization error of the current map (see Equation 6, Section 3.4) and stores it in the AQE property. Next is the CalculateUMatrix() function. This function iterates through each cell in the map and calculates the average distance between that cell and its four neighboring cells (see Equation 5, Section 3.2).

The last and most important function in the map class is TrainMap(). This function runs a single iteration of the self-organizing map algorithm each time it is called. The first step is to randomly select an instance from the training data. The selection is random so that the map does not become biased toward a subsection of the data. The .NET random class is used to generate the random seed for selection. After the selection is complete, the next step is to find the best matching unit to the selected training data. The following code snippet shows the logic for the BestMatchingUnit() function: This code block first declares and initializes a new cell called BMU. It then iterates through the map and calculates the distance of each cell in the map from the training data vector. Whenever it finds a new shortest distance, it sets the BMU cell equal to the cell in the map that is closest to the training data. Once every cell in the map has been looked at, the function returns the BMU cell.

The next step in the training process is to calculate the neighborhood radius and learning rate. The last step is also the most critical and represents the defining feature of the self-organizing map algorithm. This is where the BMU exerts its "influence" on the cells close to it on the map and "pulls" them closer to the input vector. Once the influence has been determined, there the weight of the current node is updated (see Equation 2, Section 2). Once this is complete for each dimension of each cell, the algorithm is finished for the current time step. This entire process will be repeated for each call to TrainMap() until the time step limit has been reached, at which point the map training is complete and the map is ready to be used. .

4.2 The Interface

As a windows forms application, the user interface for this implementation is contained in two parts: the visual design of the form and the code that handles initialization and events. Figure 1 shows the graphical form that the user interacts with. The form can be divided into two logical components – a series of input controls that allow the user to configure their map, and a series of outputs designed to show the result of their map training. Upon clicking the "Run" button, the user is provided with feedback on the map training, both upon completion and during the training. For brevity, we just list some of the main application outputs:

- AQE – The average quantization error of the trained map. This value is only displayed once training has completed.

- Blended View (tab 1) – This is the first of the three visualizations for the map. The blended view is the only one of the three that is updated during training so the user can observe the evolution of the map's weights while it is training.
- U-Matrix (tab 2) – The second map visualization which displays the grayscale unified distance matrix of a training map.
- Component Planes (tab 3) – The third map visualization which displays the first four component planes of the data. The coloring is based on the color wheel and scales from blue (low) to green (mid) to red (high).

5 Testing and Evaluation

Although visualization techniques are qualitative, it is possible to objectively and quantitatively evaluate the quality of a trained map. One method of evaluating the quality of a map is to calculate the average quantization error (see Equation 6, Section 3.4) of the map. This experiment will seek to observe how the various map parameters affect the average quantization error of a map. The self-organizing map algorithm is run on on three different data sets of varying size to record the average quantization error with different parameter configurations.

The data sets selected for this experiment were the Iris data set, the Wine data set, and the Abalone data set found in [7]. These three were selected because they are among the most popular and well-recognized data sets in the machine learning community in addition to providing a diverse range in the number of instances and attributes included in the data. The three data sets respectively contain: 150 instances of 4 floating-point attributes; 178 instances of 12 floating-point attributes and 1 integer attribute; and 4,177 instances of 8 floating-point attributes.

The parameters involved in testing were learning rate, number of iterations, number of cells in the map, and choice of function for the influence calculation. The neighborhood radius and learning rate decay calculations were statically set as exponential functions and were not included has variables. The learning rate for the experiment varied between 0.1 and 0.3, in increments of 0.1. The number of iterations was set at 1,000, 5,000, and 10,000. The map size was set at 100 (10 x 10), 400 (20 x 20), and 1,600 (40 x 40). The influence calculation was either linear or exponential. This results in 54 (3 x 3 x 3 x 2) parameter combinations on three different data sets for a total of 162 different maps.

5.1 Results

After conducting the experiment, the AQE for each of the 162 trained maps was recorded and an aggregate AQE was calculated for each of the four independent variables (learning rate, number of training iterations, map size, and influence calculation). The learning rate and influence calculation seemed to have relatively little effect on the average quantization error of the trained maps compared to map size and number of iterations. On average, a learning rate of 0.1, 0.2 and 0.3 resulted in an AQE of 13.28, 13.04, and 13.47, respectively. The difference between learning rate values is so small that it cannot be attributed to the parameter itself any more than random chance due to map initialization. Similarly, a linear influence calculation resulted in an average AQE of 14.29 while an exponential influence calculation resulted in an average AQE of 12.24. While the exponential influence calculation did result in maps with consistently lower AQE, the average difference was only 15.5%. Figures 2 and 3 show the learning rate and influence calculation results graphically.

In contrast to the previous two parameters, both the number of iterations and the size of the map seemed to have a significant influence on the average quantization error of a trained map. Running 1,000 iterations resulted in an average AQE of 17.16, while 5,000 and 10,000 iterations resulted in averages of 12.25 and 10.39, respectively. The average difference between running 1,000 iterations and 10,000 iterations was 49%, demonstrating that increasing the number of training iterations can significantly improve the quality of the map. Figure 4 shows these results graphically.

The results from increasing the map size were even more dramatic. The average AQE of a map with 100 nodes was 19.42, while maps of 400 and 1,600 nodes resulted in averages of 12.77 and 7.61, respectively. Overall there was an 87% decrease in the average quantization error of a trained map when going from 100 nodes to 1600 nodes, indicating the simply increasing the size of the map can drastically improve map quality. Figure 5 shows these results graphically.

An AQE of 1 can be thought of as a map having an average Euclidean distance (see Equation 1, Section 2) of 0.01 between an input vector and that vector's best-matching unit on the map. Since the vector values are normalized between 0.0 and 1.0, this number can also be thought of as a 1% average

difference between an input vector and its best matching unit.

6 Concluding Remarks

Based on these results, we can conclude that increasing the number of iterations of the training algorithm and increasing the number of nodes in the map have the most significant effect on the average quantization error of a map. By contrast, altering the learning rate or changing the function to calculate a node's influence on its surrounding nodes does not seem to have a great effect on the average quantization error. However, there are several limitations to these findings that should be noted. The number of iterations performed and the size of the maps used in this experiment were relatively limited, so we cannot make any assumptions about further reduction in the AQE should we continue to scale these values. It is also worth mentioning that the data points chosen for the map size and number of iterations in this experiment were exponentially increasing while the other parameters were either linearly increasing (in the case of learning rate) or binary (in the case of the influence calculation choice). When viewed in this light, the reduction in AQE caused by increasing the map size or number of training iterations does not seem to be quite as drastic. Despite this realization, the case remains that increasing either the number of iterations or number of nodes in the map seems to be the best way to reduce quantization error.

Additionally, and perhaps most importantly, it must be noted that average quantization error by itself is not a fully representative measurement of a self-organizing map and cannot be used alone to evaluate the quality or usefulness of a map. One might ultimately consider the effectiveness of a map to be in how much information about the data is revealed through visualizations or how useful it is for further data mining such as clustering or classification, neither of which is captured by average quantization error. However, with no objective mathematical evaluation available, the AQE measurement remains the most suitable way to quantitatively evaluate a map's quality.

7 References

[1] Kohonen, T. (1990). The Self-Organizing Map. *Proceedings of the IEEE, 78*(9), 1464-1480.

[2] Kohonen, T. (1995). *Self-Organizing Maps.* Springer, Berlin, Germany.

[3] Van Hulle, M. (2008). Self-Organizing Maps.

[4] Schatzmann, J. (2003). Using Self-Organizing Maps to Visualize Clusters and Trends in Multidimensional Datasets.

[5] Ultsch, A. (2003). U*-Matrix: A Tool to Visualize Clusters in High Dimensional Data. *University of Marburg Technical Report, 36*, 1-12.

[6] Kohonen, T., Hynninen, J., Kangas, J., & Laaksonen, J. (1996). SOM_PAK: The Self-Organizing Map Program Package.

[7] Asuncion, A. & Newman, D.J. (2007). UCI Machine Learning Repository

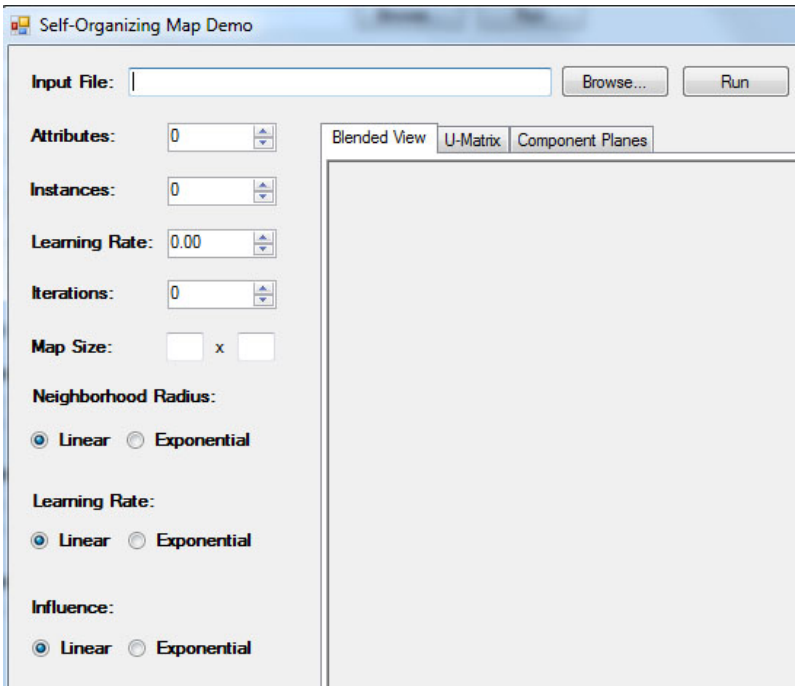


Figure 1: The graphical user interface for the SOM implementation

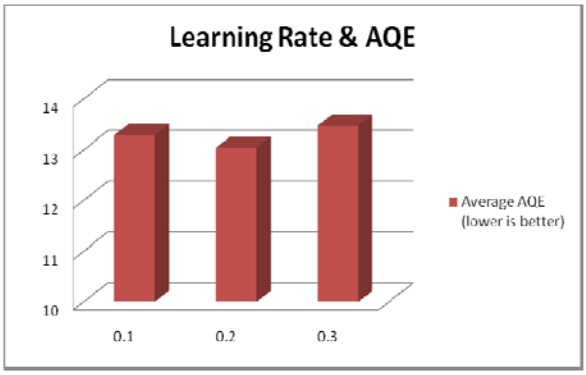


Figure 2: The average calculated AQE of the trained maps, organized by learning rate

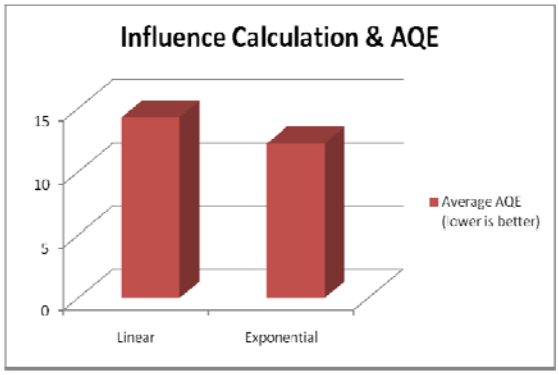


Figure 3: AQE of the trained maps, organized by the method used to calculate influence

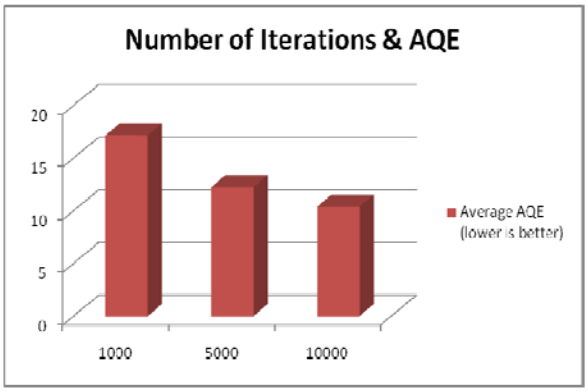


Figure 4: The average calculated AQE of the trained maps, organized by number of iterations

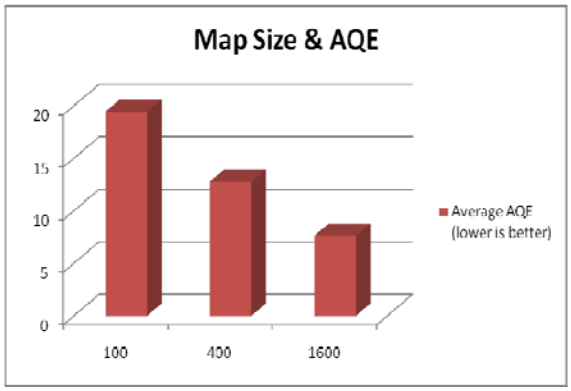


Figure 5: The average calculated AQE of the trained maps, organized by map size

SESSION

INFORMATION RETRIEVAL and DATA MINING

Chair(s)

TBA

Text Mining for Sentiment Analysis of Twitter Data

Shruti Wakade, Chandra Shekar, Kathy J. Liszka and Chien-Chung Chan

The University of Akron
Department of Computer Science
liszka@uakron.edu, chan@uakron.edu

Abstract

Text messages express the state of minds from a large population on earth. From the perspective of decision makers, this collection of messages provides a precious source of information. In this paper, we present the use of Weka data mining tools to extract useful information for classifying sentiment of tweets collected from Twitter. The results of tweet mining are represented as decision trees that can be used for judging sentiment of new tweets. We introduce a new method for preprocessing tweets for decision tree learning. We evaluate the impact of tweets containing emoticons to the classifying process. The method is applied to perform sentiment analysis from tweets related to iPhone and Microsoft. Experimental results show that decision tree classifiers out-performed naïve Bayes' algorithm.

Keywords: geometric tiling, minimal covering sets, wireless sensor networks

1. Introduction

Billions of dollars are spent worldwide each year on market analysis. Data-driven decisions are a powerful and necessary method of conducting business. Imagine how useful it would be for a company to know how its products are viewed in the market or how a political candidate could leverage their public image in their campaign, without surveying people directly. One way to accomplish this is by collecting public sentiment on Internet microblogging sites such as Twitter[1], Tumblr[2], Plurk[3], Pownce[4], and Jaiku[5]. These are the top five social networking forums that provide a quick and easy means for people to express themselves while creating a valuable pool of data for those who are interested in those opinions. Messages that users create are saved in their personal profile and forwarded to others in their circle of friends. The information may be kept private among the list, or made public and unrestricted.

Opinion mining, sentiment analysis, and subjectivity analysis are related fields sharing common goals of developing and applying computational techniques to process collections of opinionated texts or reviews. Other research goals are to generate heuristics or tools that can be used to classify, rank, or summarize sentiments toward certain objects, events, or topics. For example, these tools can be used to determine a thumbs up or thumbs down vote for specific movies from their reviews, or to predict in-favor or in-worse of certain products or events.

In this paper, we look specifically at Twitter data, called tweets, to perform clustering and sentiment analysis. Tweets are limited to 140 characters. Figure 1 shows an actual tweet taken from Twitter. This type of cyber-communication is commonly called microblogging. Sentiment analysis is a field of research that determines if there is a favorable or non-favorable reaction in text.

Figure 1. Example tweet.

Our approach is to use the Weka1 data mining software with a positive and negative word set and compare it to a second word set provided by Twitter. We are interested in the impact of emoticons added to both of these sets.

In section two, we discuss previous research in the field of sentiment analysis on text. Section three presents the problem statement and setup. In section four, we describe the preprocessing steps performed on the data and the

[1] http://twitter.com
[2] http://www.tumbl.com
[3] http://www.plurk.com
[4] http://pownce.com
[5] http://jaiku.com

feature selection used. Section five presents the experiments. Section six contains discussion of the results and we conclude in section seven.

2. Previous Work

There is a small, but growing body of research in specifically opinion mining from microblogging data. Kim et al. give a compelling case for using Twitter lists for a corpus in sentiment analysis2. In this context, lists are groups of people who share a common interest such as music. They show that even though tweets are brief, they contain enough information to express identifiable characteristics, interests and sentiments.

The seminal work by Pang et al. shows that machine learning is a viable tool for sentiment analysis using movie reviews for a corpus3. They apply three standard machine learning algorithms; Naïve Bayes, maximum entropy (MaxEnt), and support vector machines (SVMs). Their positive and negative word lists were relatively small, from five to eleven in different experiments, but nonetheless, the results are good. More notable, they bring to light the difficulty of the task compared to topic based classification.

The work in Go et al. is very similar to Pang in using the same three classifiers, but microblogging data from Twitter is used as opposed to the longer text movie reviews4. The results are remarkably similar, showing promise that applying these tools for sentiment analysis cross the boundaries from longer text blocks to the 140 characters restricted tweets. The research in this paper excludes neutral sentiments from the corpora. Only positive and negative tweets are collected, mined through queries in the Twitter search utility using common emoticons. Once collected, the emoticons are removed from the tweets before training with the classifiers. Manually collected test data retains emoticons, if present.

Pak and Paroubek5 collect data from Twitter, filter it and then classify as positive or negative by the use of popular emoticons (smiley faces, sad faces, and variations). Neutral tweets are collected from newspaper accounts to round out the corpora. An analysis indicates the distribution of word frequencies in the collection is normal. They apply a Naïve Bayes classifier to test the posts. Their best results are those experiments using bigrams. This is contrary to the findings of Pang, but may easily be explained by the very nature of the differing corpora. Movies reviews may contain more words and users may take more time to think about their post where tweeters tend to give lightening quick, brief snapshots of a thought sent from a cell phone or other small device. In fact, one very interesting observation that this paper makes is the amount of slang used and frequent misspellings in tweets. This may have minor effects on any opinion analysis applied to microblogging data.

Read performs sentiment analysis on Usenet group data and movie reviews. He uses the Naïve Bayes and SVM classifiers6. His corpus is created using emoticons to identify positive and negative texts. No neutral or objective texts are included in either the training or testing data sets. Read also looks at topic, domain, and temporal dependency classifications.

To summarize, research parameters tend to be grouped as follows:

- Classifier used
  - Naïve Bayes
  - Maximum Entropy
  - Support Vector Machine
- Text blocks versus microblogging data
- Positive/negative word list source and size
- Use of neutral/objective data
  - In the training data set
  - In the testing data set
- Use of emoticons
  - In the training data set
  - In the testing data set
- Use of unigrams, bigrams, or both
- Use of word presence versus word frequency

3. Problem Formulation

Sentiment analysis can be viewed as an application of text categorization, which dates back to the work on probabilistic text classification by Maron7. The main task of text classification is how to label texts with a predefined set of categories. Text categorization has been applied in other areas such as document indexing, document filtering, word sense disambiguation, etc. as surveyed in Sebastiani8. One of the central issues in text classification is how to represent the content of a text in order to facilitate an effective classification. From researches in information retrieval systems, one of the most popular and successful method is to represent a text by the collection of terms appear in it. The similarity between documents is defined by using the term frequency inverse document frequency (tfidf) measure9. In this approach, the terms or features used to represent a text is determined by taking the union of all terms that appear in the collection of texts used to derive the classifier. This usually results in a large number of features. Therefore, dimensionality reduction is a related issue that needs to be addressed.

The problem we consider in this paper is as follows. Given a collection of tweets related to a specific subject,

how do we come up with a classifier for labeling sentiment of new tweets as positive, negative, or neutral? We start by collecting related tweets using a query containing words or phrase denoting the subject of interest. Since tweets may belong to multiple subjects, the inclusion of a tweet to a specific subject is not necessarily certain. In this work, we do not consider a fuzzy membership.

In order to apply data mining tools to generate a classifier, we need to determine a list of features to represent tweets and assign a sentiment label to each tweet. Instead of using all terms that appear in the collected tweets, we have adopted a list of positive and negative words together with one where a positive emoticon is present and one where a negative emoticon is present to form the list of features. This is, in general, a much smaller set of features than using the unigram representation. We use three values for sentiment determined by combining the sentiment values derived from the following two factors:

(1) The frequency counts of positive and negative words.
(2) The presence of a positive or negative emoticon.

If the count of positive words is greater than the negative words, then factor (1) has value 1, else its value is -1, and it has value 0 for a tie. For factor (2), it has value 1 when only a positive emoticon is present, its value is -1 when only a negative emoticon is present, and it has value 0, otherwise. The final sentiment value for a tweet is determined by summing up the values of factors (1) and (2), and then it is mapped into one of the three possible values: positive, negative, or neutral. Table 1 contains example of each sentiment for the iPhone.

Table 1. Example iPhone-related tweets for each sentiment.

Sentiment	Tweet
positive	iPhone junkie lots talk i'm :)
negative	Anyone else frustrated MMS experience iphone? Logging slow buggy ATT website... Seems un-apple like.
neutral	Ok help here, buy phone, choices are: G1, iPhone, BB Storm, BB Bold. Chime

We use the Weka data mining program J4810 to generate a decision tree from the labeled training set. A decision tree is a symbolic classifier with two advantages: first, it can further reduce the features to be included in the tree and second, the tree structure can provide a different form of summary for sentiments derived from the training set.

4. Methodology for Sentiment Classification

The following steps were applied for text mining Twitter data for our sentiment analysis.

4.1 Data collection

We used a publicly available dataset for our sample space, provided for research purposes under Creative Commons license from Choudhury[11]. This data set contains more than 10.5 million tweets collected from over 200,000 users in the time period from 2006 through 2009. As subjects of interest, we use ñiPhoneò and ñMicrosoftò as query terms to retrieve tweets from the raw data. The iPhone corpus contains 18,548 related tweets. The Microsoft corpus consists 14,547 related tweets.

4.2 Data preprocessing

We took several steps to preprocess the data to clean the tweets. First was the removal of ñstop wordsò. These are words commonly filtered out when doing any type of text processing. In our data, we mainly removed prepositions and pronouns along with words such as *been, have*, *is*, *being*, and so forth. They can easily be removed without affecting the sentiment of the message as they do not convey any positive or negative meaning.

Itôs common to find URLs in tweets, as people often share interesting links with friends. The next preprocessing task was to identify hyperlinks in the text and replace them with the tag URL. Symbols were also removed except for those that make up the set of emoticons listed in Table 2.

Stemming is a process of reducing a word to its root form. For example, the set of words *read*, *reader*, *readers*, and *reading* all reduce to the root word *read*. We used the Snowball stemmer available as part of the Weka[1] software.

4.3 Feature Determination

We use the following features to represent tweets in our experiments. A list of 931 positive words was downloaded from Winspiration[12]. Example words in this list are *beautiful*, *easy*, and *popular*. A list of 1838 negative words was downloaded from EQI[13], a web site with resources related to emotional intelligence. Example negative words from this list are *fragile*, *grumpy*, and *stressed*. The set of emoticons we used are listed in Table

2. Positive emoticons are collectively represented as a feature named C+, and negative emoticons are collectively represented as a feature named as C-. For comparison purposes, a set of 129 positive and 144 negative words compatible to those provided by Twitrratr[14] was downloaded from the web site. These lists contain emoticons which were removed.

Table 2. Set of emoticons.

Positive set of emoticons	Negative set of emoticons
:)	:(
:-)	:-(
:)	:(
:D	
=)	
;-)	
;)	

In addition, we looked at the frequency distribution of sentiment words among the subject-related tweets. Many words have a frequency count that is less than two. Therefore, we apply a threshold of two to further reduce the features. As a result, the word list from EQI and Winspiration has been reduced from 2769 to 59 words, and the list from Twitrratr has been reduced from 273 to 30 words.

4.4 Sentiment labeling

We have created four training sets with combinations of two sets of sentiment words (EQI and Winspiration as one set, Twitrratr as the other) and inclusion or exclusion of emoticons. Training tweets are labeled by using a Java program that implements the labeling strategy described in Section 3.

5. Experimental Results

We used the Weka data mining tools for our experiments. For each of the four combinations, we create an independent testing set by randomly selecting 20% of the labeled tweets collected. The remaining 80% is used for creating classifiers using Wekaôs J48 and Naïve Bayes algorithms. The validation is done by 10-fold cross-validation. Default parameters are used for both learning algorithms. The experimental results for iPhone-related tweets are shown in Table 3. The first training set, denoted by T1-1, uses 59 out of 2769 words downloaded from Refs. 11ï 12 as its features.

Features used in the second training set, denoted as T1-2, consist of those in T1-1 plus the two emoticon categories. Similarly, features of the third training set T2-1 consist of only the Twitter compatible word list downloaded from Ref. 13 using 30 out 273 words. Similarly, the fourth training set T2-2 includes the two emoticon categories. The values of the receiver operating characteristic (ROC) areas are all excellent, and most of the F-measures are excellent, as well, as shown in Table 3. In this case, the table shows that the decision tree based algorithm J48 outperforms the Naïve Bayesô algorithm. In addition, the use of the emoticon categories as features has a negative effect on J48 learning, while they provide a slight improvement for Naïve Bayesô learning. The use of a large feature set has a negative impact on the Bayesô algorithm, but it seems to have no impact on J48.

Table 3. Performance measures for iPhone-related tweets analysis.

	Accuracy		F-Measure		ROC Area	
	J48	NB	J48	NB	J48	NB
T1-1	98.05	84.73	0.98	0.84	0.98	0.996
T1-2 (Emoticons)	97.87	85.22	0.98	0.84	0.98	0.995
T2-1	98.03	95.19	0.98	0.95	0.93	0.999
T2-2 (Emoticons)	98.03	95.41	0.98	0.95	0.94	0.999

The experimental results on Microsoft-related tweets are shown in Table 4. We have similar results as in the case of iPhone-related tweets. The J48 algorithm has out-performed the Naïve Bayesô algorithm in all cases. Again, the use of emoticons as features does not improve performance. Instead we see a slight negative impact in all cases.

Table 4. Performance measures for Microsoft-related tweets analysis.

	Accuracy		F-Measure		ROC Area	
	J48	NB	J48	NB	J48	NB
T1-1	97.56	85.61	0.98	0.84	0.964	0.998
T1-2 (Emoticons)	97.49	84.96	0.97	0.83	0.964	0.998
T2-1	97.62	95.94	0.97	0.95	0.859	1
T2-2 (Emoticons)	97.56	95.87	0.97	0.96	0.87	0.951

6. Discussion

The use of Internet slang must be addressed in any work involving microblogging data. The original motivation for users to create these abbreviations was to reduce keystrokes. Texting on cell phones made this form of writing even more pervasive. In some cases, this has grown into social cultures with different dialects (ex., leet, netspeak, chatspeak) rather than a timesaving utility. In our case, we observe that the words or phrases used in tweets may include many of these abbreviated words such as abt (about), afaik (as far as I know), alol (actual laugh out loud), and so forth. This may cause missed matches with words or phrases that appear on the positive and negative word list. To evaluate the impact of irregular expressions in tweets to our strategy of tweet labeling, we have compiled our own list of 500 abbreviated words by personal observation and various web sites. We observed that the overlap is small between this list and the positive and negative word lists used in our experiments. Therefore, the impact is minimal, which is confirmed by our experiments on the iPhone-related tweets where the hit rate of positive words versus negative words remains quite similar with and without substitutions of abbreviated words or phrases. Thus, it does not affect the result of labeling tweets based on a sentiment word list. However, the excessive amount of abbreviated words in tweets may need to be dealt with in different types of tweet analysis.

We also note that some emoticons may be neutral, for example ñ(_/)ò indicating bunny ears or ñ0w0ò meaning non-decript. We do not include these or use them as indicators of a neutral tweet. This is a possible addition to future work on tweet sentiment analysis since microblogging use and strategies are constantly evolving. We speculate on the high accuracies obtained by using the decision tree approach for classifying tweets in contrast to previous results of using Naïve Bayes or Support Vector Machine (SVM) classifiers based on different feature representation schemes of tweets. There are three possible factors:

(1) We use single subject-related tweets in our experiments for training J48.

(2) We use three values for sentiment: positive, neutral, and negative.

(3) We use sentiment words as features to represent tweets, thus reducing the impact of ñthe curse of dimensionalityò.

From our experiments, we observe that there are a large number of tweets which do not contain any sentiment words. Therefore, they are classified as neutral in our strategy. This indicates the importance of including a neutral label in sentiment analysis. The high performance obtained by J48 in classifying single subject-related tweets may suggest that the integration of document filtering techniques, described in Refs. 15, 16, and 17. This may lead to the development of even more effective systems for tweet analysis. A collection of tweets can be sorted into different categories or subjects by first applying document filtering algorithms, followed by applying single subject-related tweet analysis.

The use of sentiment words as features for representing tweets seems to be quite effective from our experiments. It is reasonable to think that the list we used happens to contain a large enough number of typical sentiment words. Thus, the availability of an effective list of words is an important factor for our approach to be successful.

It is possible that our approach can be further enhanced by integrating more sophisticated feature selection functions such as those taking into account local context18, using DIA association factor19, making use of distribution of multi-words20, or considering different similarity measures21. In addition to decision tree learning programs, there are other data mining and knowledge discovery tools22, 23 which may be used to generate and present results of tweet analysis.

7. Conclusions

In the paper, we have presented the process of applying Weka data mining tools to generate decision trees for classifying sentiment of tweets. We introduced the idea of using a list of sentiment words plus emoticons as features to represent and to label tweets for training data. We also include a neutral classification of tweets in our corpus. Experiments on iPhone and Microsoft related tweets show that decision tree classifiers out-perform naïve Bayesô ones using our approach. In addition, it appears that including emoticons as features has slightly negative impacts on the performance of decision tree based classifiers. The impact of the naïve Bayesô classifiers is mixed.

Our experiments also show that dimension reduction is critical to the performance of naïve Bayesô classifiers. Based on our approach and experimental results, we observe that the integration of document filtering and document indexing techniques with our approach may provide one viable way to the development of effective systems for tweets analysis. Our future work includes

application of our approach to tweet analysis based on different data mining tools.

8. References

[1] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. Witten, The WEKA Data Mining Software: An Update, SIGKDD Explorations, Vol. 11, No. 1, 2009.

[2] D. Kim, Y. Jo, I-C. Moon, and A. Oh, Analysis of Twitter Lists as a Potential Source for Discovering Latent Characteristics of Users, Workshop on Microblogging at the ACM Conference on Human Factors in Computer Systems (CHI 2010).

[3] B. Pang, L. Lee, and S. Vaithyanathan, Thumbs up? Sentiment Classification using Machine Learning Techniques, Proc. Of the Conf. on Empirical Methods in Natural Language Processing (EMNLP), July 2002, pp. 79-86.

[4] A. Go, R. Bhayani, and L. Huang, Twitter Sentiment Classification using Distant Supervision, Proc. of the 4th International Conf. on Computer and Information Technology (CIT2004), pp. 1147-1152.

[5] A. Pak and P. Paroubek, Twitter as a Corpus for Sentiment Analysis and Opinion Mining, Proc. of the Seventh Conf. on International Language Resources and Evaluation (LREC'10), May 2010.

[6] J. Read, Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification, Proc. of ACL-05, 43rd Meeting of the Association for Computational Linguistics, 2005.

[7] M. Maron, Automatic Indexing: an Experimental Inquiry. J. Assoc. Comput. Mach. 8, 3, 404-417, 1961.

[8] F. Sebastiani, Machine Learning in Automated Text Categorization. ACM Computing Surveys, Vol. 34, No. 1, 1-47, March 2002.

[9] G. Salton, A. Wong, and C. Yang, A Vector Space Model for Automatic Indexing. Communication of ACM 18, 11, 613-620, 1975.

[10] I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. 2nd edition, Morgan Kaufman, 2005. ISBN 0120884070, 9780120884070.

[11] M. D. Choudhury, Y.-R. Lin, H. Sundaram, K. S. Candan, L. Xie, and A. Kelliher, How Does the Sampling Strategy Impact the Discovery of Information Diffusion in Social Media? Proc. of the 4th Int'l AAAI Conference on Weblogs and Social Media, George Washington University, Washington, DC, May 23-26, 2010.

[12] Positive words download: http://www.winspiration.co.uk/positive.htm

[13] Negative words download: http://eqi.org/fw_neg.htm

[14] Twitter compatible positive and negative word list: http://www.twitrratr.com

[15] N. J. Belkin and W. B. Croft, Information filtering and information retrieval: two sides of the same coin? Communication of ACM 35, 12, 29-38, 1992.

[16] D. D. Lewis, The TREC-4 filtering track: description and analysis. Proceedings of TREC-4, the 4th Text Retrieval Conference, Gaithersburg, MD, 165-180, (1995).

[17] Y.-H. Kim, S.-Y. Hahn, and B.-T. Zhang, Text filtering by boosting naïve Bayes classifiers. Proceedings of SIGIR-00, 23rd ACM International Conf. on Research and Development in Information Retrival, Athens, Greece, 168-175, (2000).

[18] T.J. Siddiqui and U. S. Tiwary, Utilizing local context for effective information retrieval, International Journal of Information Technology and Decision Making, Vol. 7, Issue: 1, 5-21, (2008), DOI No: 10.1142/S0219622008002788.

[19] N. Fuhr and C. Buckley, A probabilistic learning approach for document indexing, ACM Transactions on Information Systems, 9, 3, 223-248, (1991).

[20] W. Zhang, T.Yoshida, and X. Tang, Disbribution of multi-words in Chinese and English documents, International Journal of Information Technology and Decision Making, Vol. 8, Issue: 2, 249-265, (2009), DOI No: 10.1142/S0219622009003399.

[21] E. Atlam, A new approach for text similarity using articles, International Journal of Information Technology and Decision Making, Vol. 7, Issue: 1, 23-34, (2008), DOI No: 10.1142/S021962200800279X.

[22] Y. Peng, G. Kou, Y. Shi, and Z. Chen, A descriptive framework for the field of data mining and knowledge discovery, International Journal of Information Technology and Decision Making, Vol. 7, Issue: 4, 639-682, (2008), DOI No: 10.1142/S0219622008003204.

[23] Q. Zhang and R. Segall, Web mining: a survey of current research, techniques, and software, International Journal of Information Technology and Decision Making, Vol. 7, Issue: 4, 683-720, (2008), DOI No: 10.1142/S0219622008003150.

Application of Bioinformatics and Data Mining in Cancer Prediction

Wafa Mokharrak, Nedhal Al Khalaf, Tom Altman
Department of Computer Science and Engineering, University of Colorado Denver, Denver, Colorado, United States of America

Abstract- *Computer-aided cancer prediction and risk assessment has become a very useful tool and is starting to be taken seriously by the medical community. Advanced bioinformatics and data mining techniques are used extensively to assist in predicting the chances of an individual patient's cancer occurrence as well as the population cancer rates in general. These techniques rely heavily on analyzing and comparing genetic and medical datasets, as well as environment-based and other factors. We have developed an expert system called the cancer predictor calculator (CPC) which predicts two specific cancer risks for women. Specifically, the CPC estimates the risk for the breast and ovarian cancers by examining a number of user-provided genetic and non-genetic factors. The expert system was validated by comparing its predicted results with the patients' prior medical information (and the subsequent outcomes) contained in actual health history databases and other well-known sources (e.g., see [14-17]).*

Keywords: Cancer prediction calculator, breast cancer, ovarian cancer, bioinformatics, data mining, health IT

1 Introduction

Many existing information and/or knowledge-based technologies have aided humans in predicting abnormal genes and other factors that lead to diseases. Cancer is considered to be the number one fatal genetic disease.

Contrary to popular opinion, the excessive retention and/or compilation of the immense amounts of biological data have turned its analysis into a very difficult and complex undertaking. Even with the emergence of bioinformatics and data mining, and combining biology, computer science, information technology, statistics, and mathematics, the problem of efficient knowledge extraction is increasingly becoming more difficult. One of the primary purposes of bioinformatics is to clarify the biological processes that depend on hereditary resources. Data mining has the capability to detect hidden useful patterns between datasets objects and to use them as predictors. Consequently, any interaction between bioinformatics and data mining can only serve to improve their usefulness and the overall outcomes.

We have integrated both to construct the CPC, which using huge biological datasets incorporates the patient-provided information to derive its prediction results. This software predicts the patient's cancer risk/percentage and classifies this risk into four categories (no risk, low, medium, or high). The importance and applicability of this software comes from a potential early warning, which could lead to a discovery of cancer in its early stages, when it can usually be treated with much higher success, and/or to take precautions and provide additional knowledge to the patient about any future risks.

The rest of this paper is organized as follows: Section 2 addresses cancer prediction techniques and how they are being applied by the CPC. A short discussion of Gail algorithm that predicts breast cancer is presented in Section 3. Our contributions, results and conclusions are given in Sections 4 and 5.

2 Cancer prediction techniques

Cancer prediction is certainly a very complex and nondeterministic endeavor. Estimating the probability of cancer occurrences in patients requires that many factors (both genetic and non-genetic) are evaluated and properly weighted according to their significance and/or other (context sensitive) contribution factors. Some of the approaches in this research include:

2.1 Decision tree

A *decision tree* (DT) is a graphical representation of sequential decisions and of all of the possible paths resulting in predictions or classes into which data objects can be classified. Decision trees are considered to be a powerful tool for classification and prediction. They are widely used in knowledge systems and smart databases [9, 10].

The CPC uses DTs in order to predict whether a gene is normal or abnormal. Here, the decision nodes are the predictors and the leaves are the terminal predictions or the target. A decision tree structure allows us to classify patients into several risk levels. For each patient, it can be used to identify the likely group membership and to show the relationship

and/or effects between the identified cancer causing factors.

We have applied DTs to the CPC as shown in Figure 1. The predictors shown represent some of the identified factors that cause and/or contribute to the development of the disease. The branches that come out from the predictors represent all the possible outcomes. The terminal nodes show the risk group determination/membership.

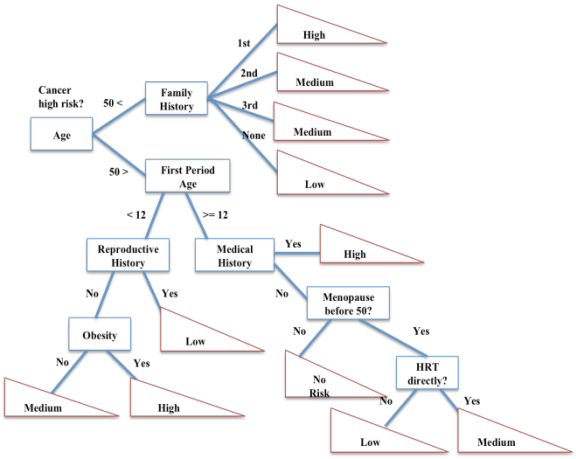


Figure 1: DT from the CPC [8, 12]

2.2 Network approach for cancer prediction

The idea behind *molecular networks* is to predict the presence of the true gene(s), which are known to be causing a specific disease or cancer. It assumes that there is a number of genes *n* that have genetic information and each of which is considered to be a potential candidate. In there is no genetic information available, the all-out human genome can represent the candidate list. The linkage after that is performed by leading to more than one gene candidate interval, where the genes are ranked according to their association with an identified specific disease or cancer. Such intervals may contain several hundred genes which, if defective, may eventually cause genetically based cancer or other disease(s).

Next, the candidate genes and all other related factors are mapped to a human gene/protein network. Then, each candidate gene is scored by a scheme based on other factors and its relative position (genes are well-ordered linearly which means homogeneous genes are located at the same relative position in their respective human gene/protein network). Finally, the genes are ranked based on their scores from top (highest) to down (lowest), where the top gene is the one that is most likely to cause the disease. This score is then confirmed by cross-validation (one trend among many, assesses the validation of training set in a scheme on unknown genes) with a known disease-genes relationship [13].

When applying this approach on ovarian cancer in the CPC, there will be many candidate genes that have genetic information about ovarian cancer and other diseases. The candidate genes (BRCA1, BRCA2, HNPCC, MLH1, MSH2, and MSH6) with all other disease's factors (e.g., personal, reproduction, genetic, hormonic, operation, birth control, body, nutrition, and lifestyle factors) are mapped. After that, each candidate gene is scored by a scheme relying on their relative position in the network and on other factors. Then, all candidate genes are ranked based on their score. For example, for BRCA1 its score is 40%, BRCA2 = 20%, MLH1, MSH2, and MSH6 = 12%, and HNPCC = 1.2%. The highest score of the candidate genes, i.e., BRCA1, will turn out to be the main cause of the ovarian cancer.

2.3 Support vector machine

Another popular technique is the *support vector machine* (SVM). Its goal is to get a hyperplane, which isolates the data associated with two different classes, and increases the lower margin that connects the hyperplane with each individual data class in the Euclidean space, even if the data are not breakable. The purpose of the increment is to select the most powerful hyperplane. If the data are not breakable, SVM tries to keep the total error, i.e., the distance from the hyperplane to the incorrectly classified samples, below a certain user-defined threshold [6].

In the CPC a linear separator is applied using statistical and standard programming techniques to make a determination if a cancer risk exists or not, as shown in Figure 2.

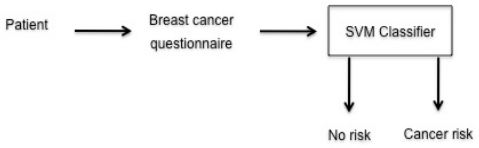


Figure 2: SVM Classifier [12]

2.4 Hidden Markov Model

The *hidden Markov model* (HMM) uses a statistical representation of multiple sequence alignments. It assumes that the probability of availability of a base at a specific position is based on the *k* previous nucleotides, where *k* represents the order in Markov model. It can be visualized as a

finite state machine (FSM) with emitting states. An FSM operates via a succession of state transition moves, which are defined by the machine's *state transition function* or *relation* for the deterministic and nondeterministic FSMs, respectively. It then presents an output indicating whether the machine has approached a certain state or while it is moving from one state to another. The HMM model topology is stable, which means the transitions and emissions probabilities are estimated values [11]. The model is expressed statistically by the expression: P(X/*k*), where the X is a given base: A, T, G, or C, and *k* is the location of previous nucleotide

The HMM has been applied on the CPC for breast cancer tumor risk assessment, as shown in Figure 3.

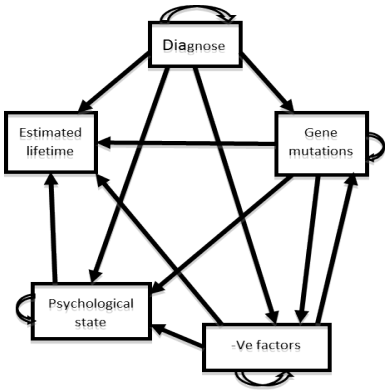


Figure 3: The breast cancer HMM [12]

2.5 Bayesian Networks

The *Bayesian networks* (BNs) are characterized as being members of the probabilistic graphical model family. Nodes in the graph represent random variables, and the edges represent the probabilistic dependencies between random variables. They are given as either a conditional probability function or via a table.

A BN consists of a directed acyclic graph (DAG) that represents casual relationships between arbitrary variables and parameters. Combing both graphical structure and conditional probability illustrates the full power of a BN probabilistic model.

BNs are frequently used in bioinformatics for the task of integrating various genes prediction systems. They allow finding a closest corresponding network to the existing training set of independent parameters. This process can be obtained based on statistical function, called the *scoring function*, which finds the optimal network by evaluating each network with consideration to the training set [1, 9, 11].

The application of BNs on the CPC, which was achieved by combing both graphical structure and conditional probabilities, is shown in Figure 4.

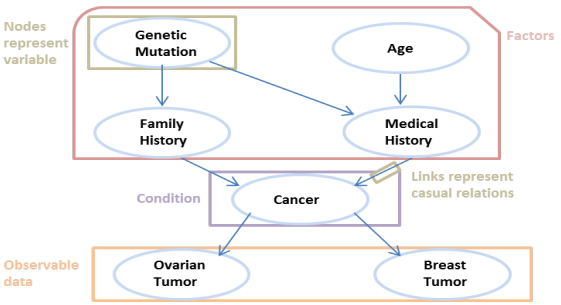


Figure 4: The BN used with the CPC [12]

In the DAG shown in Figure 4, *genetic mutation* and *age* represent the independent variables. When applying independent condition formulas on them, we get:

P (Genetic mutation and Age) = P (Age) * P (Genetic mutation).

Here, P (Age) stands for probability of a woman getting cancer at her specific age, and P(Genetic mutation) stands for probability of a woman getting cancer because of having any gene mutation.

Also, the cancer variable occurs dependently by the patient's medical history, that is, it is derived from the independent variables genetic mutation and age. Therefore, applying the condition formulas, we get:

P (Cancer | Genetic mutation, Age, Medical history) = P (Genetic mutation, Age, Medical history and Cancer) / P (Genetic mutation, Age, Medical history).

The probability of cancer can be determined by family history or/and the patient's medical history. In contrast, ovarian tumor and breast tumor are independent to give cancer (breast or ovarian, respectively), as shown by the following formula:

P (Breast tumor | Ovarian tumor) = P (Breast tumor).

2.6 Decision rules

A group of rules that can be extracted either from a decision tree or a dataset in order to predict to which class a dataset's record belongs are called *decision rules* (DRs). They usually present all the possible predictions and decisions that are applicable in a specific situation. The DRs approach is useful in many other fields besides disease prediction and clinical diagnoses.

In the clinical field, the DRs used for gene prediction are usually called *clinical prediction rules* or *clinical decision rules*. These rules usually include predictors that are variable, extracted from some patient history such as disease characteristics, patient characteristics, or patients' physical examinations.

Furthermore, genetic and non-genetic professionals in cancer prediction can apply some these rules in predicting cancer in their patients [3].

Applying DRs to the CPC results are shown in Table 1. The *predictors* of a disease or cancer are those which significantly increase its calculated risk. The results from these predictors will give a good indication to which class of risk a patient belongs (e.g., low, medium, high).

Table1: Prediction Decision Rules from the CPC [8]

Prediction rules Preconditions	Ovarian Cancer Predicted Result
If (talcum powder="Y") ∧ (Fertility_Drugs="Y") →	Medium Risk
If (Age>=80) ∧ (Reproductive_History = "Y") →	Low Risk
If (Tubal_Ligation = "Y") ∧ ((Exercise = "Y") ∨ (Family_History = "Y")) →	No Risk
If (Period_Age<12) ∧ (Reproductive_History ="Y") →	Low risk
If (HRT = "Y") ∧ (Reproductive_History ="Y") →	Low risk
If (Ovarian_Removed = "Y") →	No Risk
If (Period_Age<12) ∧ (HRT ="Y") →	Medium Risk
If (Overweight = "Y") ∧ (Endomtrosis_Issue = "Y") →	Medium Risk
If (Smoker = "Y") ∧ (Exercise = "Y") →	No Risk
If (Gene_Mutation= "Y") ∧ (Fertility_Drugs="Y") →	Medium Risk

2.7 Association rules

Association rule (ARs) is an approach in discovering relations among unrelated data in a database or data repositories, and detecting patterns of relationships among the attributes of the dataset. In addition, this approach is very effective in many fields such clinical diagnosis, and analyzing the associations between rules for disease prediction.

Association rules can be used to predict cancer by making use of the patient's genetic and non-genetic information. Association rules in the process of predicting cancer will detect the genetic factors that are related to each other [4].

An example of applying this approach to the CPC is shown below. The factors that are related to each other are predictors for the disease. After analyzing the relationships between the predictors, determining the disease risk will be easily based on the analysis. For example, some of the association rules that can be extracted from CPC are the below. In each association rule, the family history degree is expressed as a variable (Family_History) that takes one of four values (0: none, 1: first degree, 2: second degree, 3: third degree).

(Race = "White") ∧ (Family_History = "1") ∧ (Num_Affected_Relatives > 1) → high risk.

(BRCA1_Mutation = "Yes") ∧ (Family_History = "1") → high risk.

3 Gail model for breast cancer prediction

This model has been invented by Dr. Mitchell Gail at 1989, senior investigator in the Biostatistics branch of NCI's division of cancer epidemiology and genetics [5, 7, 14].The model predicts breast cancer risk based on the five predictors used in the Breast Cancer Detection Demonstration Project: age, age of first period, age at first live birth, breast biopsies, and the first degree of family history. The Gail calculator is acknowledged as the first computer program tool for breast cancer risk assessment from the NCI.

The model predicts breast cancer for a woman for the next five years and for her lifetime. It operates by taking into account the relative risks (predictors) of breast cancer and the woman's current age, which is considered the most effective relative risk indicator. The Gail algorithm has several drawbacks:

- Gail predicts only breast cancer while CPC predicts breast and ovarian cancer.
- Gail model does not take in consideration second degree relative (nieces, aunts, grandmothers), and third degree relatives (cousins) who had been diagnosed with the disease.
- Gail model does not take into account other risk factors that play a role in increasing the risk of breast cancer. These include hormone therapy, age at menopause, gene mutations, and radiation exposure.
- Gail model may underestimate the woman's risk of breast cancer since it only concentrates on five risk factors
- It does not take into account abnormal genes that are considered a significant breast cancer risk factor.

4 Contribution and discussion

In this paper, we have discussed recent research results addressing cancer prediction and, specifically, the roles of bioinformatics and data mining. Techniques from these were applied to construct the CPC expert system. Its purpose is to predict the relative risk of ovarian and breast cancer in women. The proposed techniques can be easily generalized to build any cancer predictor for women as well as men.

The risk computed by the CPC is based on factors that are known to affect the disease. Some of the factors are increasing while others are decreasing the risk. Each questionnaire (form) in the CPC addresses several of the disease's factors, where some of them are dependent on each other regardless if they are in

the same form or not. For example, the relative risk of a factor is determined after performing arithmetic multiplication, summation, and/or subtraction on its dependent factors. After the relative risk of each factor in a form is determined, we sum them together, and do the same for each form. Finally, we add the results of all the forms to get the overall predicted risk.

The CPC calculates individual's overall predicted disease lifetime risk, denoted by *R*, as the *approximate* summation (note that certain events, e.g., having children and breast-feeding, are not completely independent, thereby necessitating certain adjustments) of all of the identified relative risk factors, i.e.,

$$R \approx \sum_{i=1}^{n} \sum_{j=1}^{m} r_i^j, \qquad (1)$$

where,

i is the form number in the application,
n is the total number of forms,
j is the question number in each form,
m is the number of questions in each form,
r_i^j is the relative risk contributed by the factors from the j^{th} question of the i^{th} form.

The issues such as conditional probability and its effects on the final calculation of *R*, as well as the instances (e.g., when certain questions had not been answered) were addressed properly by the CPC, but the details are omitted here for clarity of presentation.

We had also developed a mathematical formula, which was derived from the patients' records database, that computes the contribution of breastfeeding duration, a major factor influencing (i.e., reducing) the breast cancer risk for women. This formula, confirms the well known decrease in the predicted risk of breast cancer based on the number of months a woman had breastfed after giving birth, and her overall number of children [2]. Each group of women (associated with the number of breastfeeding months) has a specific relative risk, denoted by BM. This risk is inversely proportional to the number of breastfeeding months. The same applies if a woman has given birth to more than one child. The breast cancer risk will decrease by 8 percent for each child. Hence, the CPC accounts for the breastfeeding duration and number of children as follows:

BreastFeeding = BM + (0.08 * No. of Children).

Real datasets including patients' information such as age, family history, etc, have been used to validate cancer predictor calculator. For example, WHI (Women Health Initiative) [17] databases are often used to validate breast cancer predicted risk percentage. On the other hand, SEER (Surveillance Epidemiology and End Results) [16] databases are databases used to validate ovarian cancer predicted risk assessment.

In order to determine which calculator is the most accurate in predicting the cancer risks for women; we compared our results using several real patients' database's results. We found that the CPC more accurately predicts breast cancer occurrences than its two main competitors, the Gail model and the Australian Government Calculator (AGC) [15]. Also, the CPC's predicted ovarian cancer results were more accurate than those predicted by SEER.

entered the records of 1,760 patients, all of whom eventually got cancer, from the WHI database into the two calculators (CPC and Gail). The results, shown in Table 2, indicate that the CPC was more accurate.

Table 2: Comparison between Gail and CPC

Expert System	Predicted Cancer	Percentage of correct predictions
Gail	1496	0.850
CPC	1616	0.918

The Australian Government Calculator (AGC) doesn't consider patient's medical history, which is one of the most substantial factors affecting the predicted risk. Therefore, in order to make our comparison fair, Table 3 shows the predicted results for the same patients WHI without the medical history information. Consequently, the risk levels predicted by the AGC and CPC are very close to each other and both clearly outperform the Gail model.

Table 3: Comparison between Gail, CPC, and AGC

Expert System	Predicted Cancer	Percentage of correct predictions
Gail	1002	0.569
CPC	1400	0.795
AGC	1405	0.798

Together, Tables 2 and 3 show that CPC's predicted results are more accurate than Gail's and Australian's. The reason behind this is that CPC incorporates more cancer related factors in making its risk assessment calculations. Based on our usage to the three cancer risk assessment tools (CPC, Gail calculator, and AGC), it became noticeable that each system has its strengths and weaknesses. However, these will not be discussed here.

The CPC integrates additional factors that are not considered by the Gail calculator or the AGC. These significantly contribute to CPC's predictive accuracy:

A. *Personal factors*: last age of birth, race (it includes all ethnicities).

B. *Medical history factors*: abortion, cancer disease, and Diethylstilbestrol drug usage.

C. *Lifestyle factors*: smoking, exercise, wearing bra, and night shift work.

D. *Genetic factors*: BRCA genes mutations, obesity, and extended family history.

E. *Body factors*: head circumference at birth, weight at birth, and height at birth.

F. *Environmental factors*: radiation, living environment, and working environment.

The CPC does have some disadvantages: it is overestimating certain races' risk ratios, as compared to the Gail model and other public breast cancer calculators. It is not providing the integration of more than one degree of the family history simultaneously, as is done by the AGC. Both of these shortcomings will be addressed in the future versions of the CPC.

Moreover, after evaluating the CPC model and comparing it with other well-known breast and ovarian cancer risk predictors, additional features will be incorporated into CPC's next version:

1. Ability to draw a patient's path on the decision tree in order to let the patient know the factors that most likely caused (or are likely to cause) the disease.
2. Integrate the CPC with other systems in order to give the user more than one risk ratio estimation.
3. Include other cancer types.

5 Conclusion

Bioinformatics and data mining today are challenging interdisciplinary sciences. Both fields have played a major role in improving cancer prediction and risk assessment since they do not only extend the analysis's process, but also its depth. In fact, their techniques have a promising future for improving the efficiency of the health care and in predicting cancer prior to its occurrence. The proposed CPC prediction model for breast and ovarian cancers is fully operational and available to the general public. It has been validated by applying real databases on it and it outperforms two of the more popular cancer risk assessment systems against which it was evaluated.

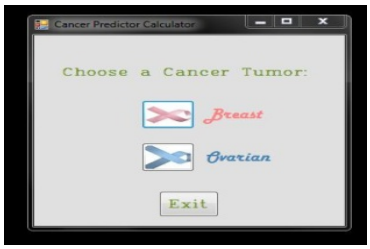
Applying proven techniques from bioinformatics and data mining, the CPC makes several practical technological contributions and enhancements, which assist with the identification of cancer prediction factors. It performs favorably when compared with state of the art systems in the area.

6 APPENDIX

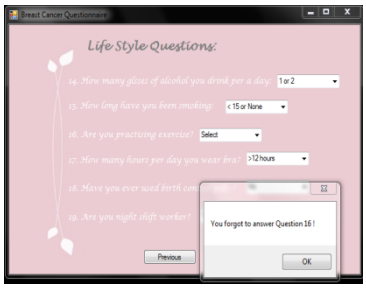
The CPC runs on Window OS. The expert system was created by Visual C#2010 Express and SQL database to design a graphical user interface and connect database tables, which contain the factors that lead to both cancers: breast and ovarian. It is at

http://www.mediafire.com/?ssbb6l8wwn7twpc

with detailed step-by-step instructions. After guiding the user through the introductory window, the CPC Menu will prompt for a choice between the breast and the ovarian cancer risk assessment, as shown below.



The user will asked to answer a number of questions in consecutive windows. In the questions' window, she can know her predicted risk by clicking the "Risk Probability" button.



The final window includes the user's the predicted risk percentage and the class of risk to which she belongs.

7 References

[1]. A. Hodges, P. Woolf, and Y. He. "Prediction of Novel Pathway Elements and Interactions Using Bayesian Networks," unpublished.

[2]. B. Marsh, "Breast-feeding reduces cancer risk," *Mailonline*.

[3]. B. Ingui, and M. Rogers. (2001, July-August). Searching for Clinical Prediction Rules in Medline. Journal of the American Medical Informatics Association. [Online]. 8(4). pp. 391–397. Available: http://www.mendeley.com/research/searching-clinical-prediction-rules-medline/

[4]. E. Louie, and T. Young. (2000). Finding Association Rules Using Fast Bit Computation: Machine-Oriented Modeling. *SpringerLink*. [Online]. 1932/2010. pp. 486−494. Available: http://www.springerlink.com/content/cc3hx3r8g09g-a99j/

[5]. J. Culver, J. Hull, E. Levy-Lahad, M. Daly, and W. Burke. (2000, March). Breast Cancer Genetics - An Overview. *GeneClinics*. [Online]. 13(1). pp. 1-12. Available: http://web.udl.es./usuaris/e-4650869/docencia/segoncicle/genclin98/malalties/Breast%20Cancer%20.pdf

[6]. L. Chuang, K. Wu, H. Chang, and C.Yang, "Support Vector Machine-based Prediction for Oral Cancer Using Four SNPs in DNA Repair Genes," in Proceedings of the International MultiConference of Engineers and Computer Scientists, Hong-Kong, 2011.

[7]. L. Newman. (2005). Breast Cancer in African-American Women. *The Oncologist Breast Cancer*. [Online]. 10. pp. 1-14. Available: http://-theoncologist.alphamedpress.org/content/10/1/1-.full.pdf+html

[8]. N. Al-Khalaf, "Cancer Prediction in Data Mining," Master thesis, Department of Engineering and Applied Science, University of Colorado Denver, Denver, CO, 2011.

[9]. P. Spirtes, C. Glymour, R. Scheines, S. Kauffman, V. Aimale, et. al, "Constructing Bayesian Network Models of Gene Expression Networks from Microarray Data," unpublished.

[10]. P. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining.* Boston, MA: Pearson Education, 2006, pp. 20-100.

[11]. S. Bandyopadhyay, U. Maulik, and D. Roy. (2008, January). Gene Identification: Classical and Computational Intelligence Approaches. *IEEE Transaction on Systems, Man, and Cybernetics.* [Online]. 38(1). pp. 55-68. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=04378438

[12]. W. Mokharrak, " Cancer Prediction in Bioinformatics Role," Master thesis, Department of Engineering and Applied Science, University of Colorado Denver, Denver, CO, 2011.

[13]. X. Wu, and S. Li, *Cancer Systems Biology*. Boca Raton, FL: Chapman & Hall/CRC Mathematical and Computational Biology Series, 2010. pp. 191-208.

[14]. Gail model. *Microcalcification Resource Site* [Online].Available: http://www.infoacademy.gr-/microcalc/index.php?option=com_content&task=view&id=20&Itemid=41

[15]. Calculate your risk. *Australian Government* [Online]. Available: http://canceraustralia.nbocc.org.au/risk/calculator.php

[16]. Datasets & Software. *National Cancer Institute* [Online]. Available: http://seer.cancer.g-ov/

[17]. Data. *Women's Health Initiative Scientific Resources Website* [Online]. Available: htt-Ps://cleo.whi.org

Situational Analysis Based on Graph Structuralization

Taketo Matsunaga[1,2], Koji Kitamura[2], Yoshifumi Nishida[2], Hiroshi Takemura[1]

[1] Department of Mechanical Engineering, Tokyo University of Science, Chiba, Japan
[2] Digital Human Research Center,
National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan

Abstract—Despite the pervasiveness of situation data such as incident reports and situation reports, we lack methods for describing and analyzing situation data. In this research, the authors propose a new system for a situational structure analysis by formulating the problem of situation mining as a problem of graph structural analysis. The proposed system consists of two basic functions: a function for graph structuralization and a function for situation data mining. The graph-structuralizing function allows users to create a semantic graph from free-description sentences about a situation. The data mining function allows users to conduct clustering, search similar situations, and visualize typical situation processes. To evaluate the effectiveness of the developed system, we analyzed 818 child-bicycle accident data.

Key Words: Situational Analysis, Graph Structural Analysis, Data Mining, Information Search

1 Introduction

In recent years, an enormous volume of situation reports related to injury has become publicly available. Knowledge creation from such a large number of reports is strongly required for a scientific approach to injury prevention, risk management, consumer product improvement, and risk communication [1][2]. Knowledge creation from serious and relatively rare accidents such as airplane crashes, plant accidents, and power plant accidents has been studied in conventional research. For example, as a pioneering work, a failure knowledge database was created and is available in Japan [3][4]. However, we still lack a good methodology for dealing with the knowledge creation of accidents that are individually relatively small in scale but very large in number, such as childhood injuries. For these types of accidents, the total scale becomes very large. According to the world report on childhood injuries published by the World Health Organization (WHO) in 2008 [5], unintentional injury is a major killer of children under the age of 18 and is responsible for approximately 950,000 deaths per year.

Finding typical situations by using situation reports, which are given as text data, and counting the frequency of each typical situation is one of the most important steps for analyzing a situation. However, it is difficult to accomplish this task by a conventional keyword search method or a text mining system. For example, when trying to find the number of situations of "a child rode a bicycle," the keyword search gives us the results of irrelevant situations such as "a mother rode a bicycle while her child was sitting in the back seat" because the keyword search system searches for text that includes "child," "rode," and "bicycle."

The second problem lies in finding the typical process of a situation. To prevent injury, we have to clarify situational structures and find the factors calling for intervention. In this paper, a situational structure indicates two kinds of structural data: the relationship among factors such as environment, consumer products, and persons, and the process of time-series change of the relationship among the factors. The time-series change means that, in general, the relationship among factors changes before the incident, during the incident, and after the incident. So we have to clarify not only the relationship among factors but also the time-series process of the relationship to intervene and control the situation so as to prevent incidents. However, no good technologies exist to support this task.

To solve these problems, this study applies a method for a graph structural analysis to a situational analysis. Technologies for a graph structural analysis [6][7] are available and have been applied to fields such as social networks [8][9], bioinformatics, and molecular structure analysis [10]. Analyzing tools and visualizing tools have also been developed [11]. If we can structuralize incident situation data as graph data, we can formulate the problem of situation mining as a problem of structural analysis.

This paper proposes a situation mining system that allows a user to find situational structures based on a method of graph structuralization, to cluster situations based on the structuralized situational data, and to visualize a typical situation process using the large amount of text data on the situation. To evaluate the effectiveness of the system, we analyzed the real data of 818 child-bicycle incidents.

2 Development of Situational Analysis System Based on Graph Structuralization

Figure 1 shows the configuration of the developed system. The developed system has two basic functions: a function for graph structuralization and a function for situation data mining. The first function, graph structuralization, enables the creation of semantic graphs from free-description sentences about an incident situation. Specifically, the system has functions for graph structuralization of situation data, management of a situation graph database, and management of a domain-specific terminology dictionary. The situational analysis allows us to conduct situation semantic search, situation clustering,

situation linkage analysis and visualization of a typical situation process.

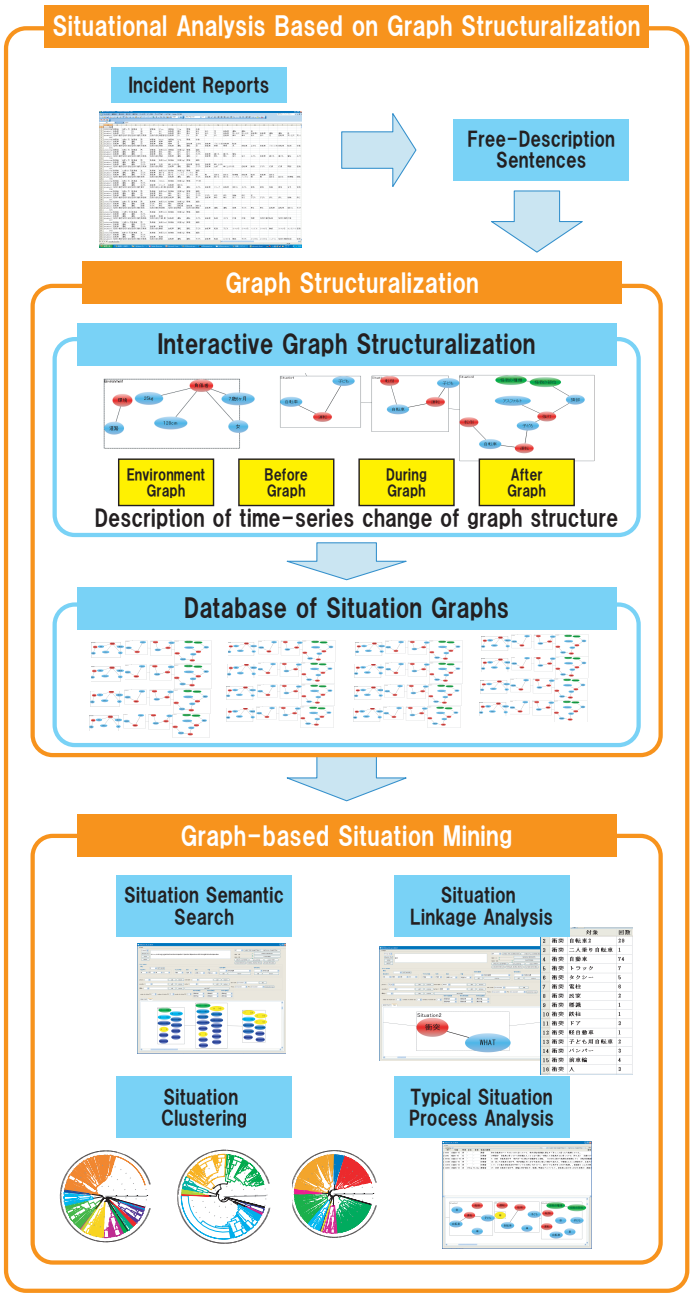


Fig. 1. Configuration of the developed situation mining system based on graph structuralization

2.1 Graph Structuralization

2.1.1 Interactive Graph Structuralization

A user can conduct graph structuralization interactively by using the developed system. We utilize MeCab [12], which is a text mining software, to support the user's graph structuralization task. For example, in the case of road accidents, the system automatically divides free-description sentences about the incident situation into individual words and categorizes them into five groups: Environment, which is used for describing attributes of an occupant and the environment, Persons, Vehicles, Things except vehicles, and Action. Secondly, it outputs nodes with labels of the dissolved words into the software GUI for creating situation graphs. This graph-editing GUI is used for describing both the relationship among components of the environment where an incident occurs and the time-series process of change of the relationship. Then, using the GUI, a user can create a graph structure by connecting the nodes. The user creates four kinds of graph structures of the situations: the environment of the situation (Environment graph), before an incident (Before graph), during an incident (During graph), and after an incident (After graph). The following are the directions for graph structuralization for accident data.

Directions for Graph Structuralization

- Rule 1. In the "Environment box," which is the GUI of the system for editing an Environment graph, the user creates the graph data that describes the attributes of the environment and the occupants. In the "Before box," the user creates the graph data that describes the situation before the incident. In the "During box," the user creates the graph data that describes the situation from the time when the incident happens to the time when the occupant is injured. In the "After box," the user creates the graph data that describes the situation after the incident.
- Rule 2. In each box, the user creates graph data such as an Environment graph, a Before Graph, a During graph, and an After graph by selecting nodes from the candidates displayed in a pop-up menu.
- Rule 3. If the situation sentence pattern is Subject + Verb, the user connects a Subject node to an Action node (a Verb node).
- Rule 4. If the situation sentence pattern is Subject + Verb + Direct Object, the user connects a Subject node to a Direct Object node through an Action node (a Verb node).
- Rule 5. If the situation sentence pattern is Subject + Verb + Indirect Object + Direct Object, the user connects a Subject node to a Direct Object node, and an Indirect Object node and a Direct Object node to an Action node (a Verb node).
- Rule 6. The user creates each graph by adding a new graph to the situation graph of the previous phase. For example, the user can create a During graph by adding an additional graph to the previously created Before graph.

2.1.2 Management of Situation Graph Database

The database function manages the graph-structuralized situation data. The user can retrieve the registered data and modify the data. This function is mainly used for the situation mining functions, described later.

2.1.3 Management of Terminology Dictionary

The text mining engine requires a dictionary of words. In the developed system, we have to create a dictionary of domain-specific terminology and classify the words into five categories, such as Environment, Persons, Vehicles, Things, and Action. The data mining function is used for supporting this task. If the system finds unknown words while the user

proceeds with graph structuralization, the system registers them into the dictionary by interactively asking the user for a suitable category for the unknown words. As the graph-structuralization task proceeds, the dictionary becomes rich and the user is able to skip the process of registering the unknown words by hand.

We also implemented a dictionary of synonyms. In this function, one representative word is selected for each group of synonyms so that we can convert every word into its representative word, if a representative word exists. This is one of the common functions of general text mining software. We expand this function for use in the situation graph database. When the system finds a new word, the system lists the candidates of synonyms for the new word by checking the words connected to the new word and the semantic connection in the graph database.

For example, if the graph database has a situation graph corresponding to the situation "a bicycle collided with a car," then in the dictionary the word "collide" is registered as a word connected to nodes with the labels "bicycle" and "car." When creating a situation graph corresponding to the new situation "a bicycle crashed with a car," which includes a new word "crash," the system recognizes that "crash" has the same semantic structure as "collide" by using the synonym function. The user can register new words by representatives from the candidates.

2.2 Situation Mining

2.2.1 Situation Semantic Search

The situation semantic search function enables us to search data considering the time-series of the situational structures. It is difficult to do this kind of search by using a simple keyword search algorithm. Users can search situation data by creating a situation graph as a search query. If users set any nodes as "necessary condition," the user will get results that not only satisfy Eq. (2), but also the nodes existing in the graph structures of the results.

In the algorithm of the situation semantic search, we can compute the similarity between two different situations by the rule that "two graphs are similar if they share many edges in common." We represent each phase of a graph-structuralized situation as G (Before graph: G_B(v,e), During graph: G_D(v,e), After graph: G_A(v,e)), and the number of edges that two different situations G_i G_j share in common as Sum($G_i \cap G_j$). The similarity among two situation graphs can be defined as follows:

$$Sim(G_i,G_j)=\frac{Sum(G_{iB}\cap G_{jB})+Sum(G_{iD}\cap G_{jD})+Sum(G_{iA}\cap G_{jA})}{Sum(G_i)}. \quad (1)$$

We can search the desired situation data by calculating the similarity expressed by Eq. (1). Specifically, by defining the following condition expressed by Eq. (2), we can find a similar situation.

$$Sim(G_i, G_j) > S_{min}, \quad (2)$$

where Sim_{min} is the minimum value of similarity that ranges from 0 to 1.

2.2.2 Linkage Analysis on Situation Graph

The linkage analysis function allows us to compute which nodes are connected to a specific node and the frequency of of connections. Using the developed software stated later, the user can set a search condition by simply connecting the specific node to a "WHAT" node. Then, by pushing a "Search WHAT" button, the system finds the nodes connected to the specific node and counts how many times the node is linked to other nodes.

2.2.3 Situation Clustering

The clustering function conducts cluster analysis by various kinds of clustering methods such as K-means, hierarchical clustering, and graph kernel. The cluster analysis is conducted separately for each phase of situations, such as Before graphs, During graphs, and After graphs. The detail of the algorithm is as follows. First, the system computes unique graph structures in each graph G' (G'$_B$(v,e), G'$_D$(v,e), G'$_A$(v,e)) by the following formulas:

$$G'_B = G_B, \quad (3)$$
$$G'_D = G_D - G_B, \quad (4)$$
$$G'_A = G_A - G_D. \quad (5)$$

For example, Fig. 2 and Fig. 3 show a situation graph for "a child rode a bicycle and fell down with the bicycle." In this case, the graph in Fig. 2 is G'$_B$(=G_B), and the graph in Fig. 3, in which the nodes and edges colored gray are expressed to be the common graph with G_B, is G'$_D$. The similarity matrix is expressed by Eq. (6), in which the suffix "D" indicates the matrix for the During graph.

$$\mathbf{Msim_D} = \begin{pmatrix} 1 & Sim_{D12} & ... & Sim_{D1n} \\ Sim_{D21} & 1 & ... & ... \\ ... & ... & 1 & ... \\ Sim_{Dn1} & ... & ... & 1 \end{pmatrix}, \quad (6)$$

where Sim_{Dij} indicates the following value:

$$Sim_{Dij} = \frac{Sum(G'_{Di} \cap G'_{Dj})}{Sum(G'_{Di})}. \quad (7)$$

The distance matrix $Mdist_D$ for this clustering is the same as the average of $Msim_D$ and the transposed $Msim_D$.

$$\mathbf{Mdist_D} = \frac{\mathbf{Msim_D} + \mathbf{Msim_D^T}}{2}. \quad (8)$$

2.2.4 Visualization of a Typical Situation Process

The visualization function visualizes the process of typical situations by using the results of clustering. First, the system conducts a cluster analysis separately for each phase of the graphs (Before graph, During graph, and After graph). In this procedure, typical situations for each phase are clarified and displayed as nodes in the developed software. Second, the system analyzes the typical connections among different

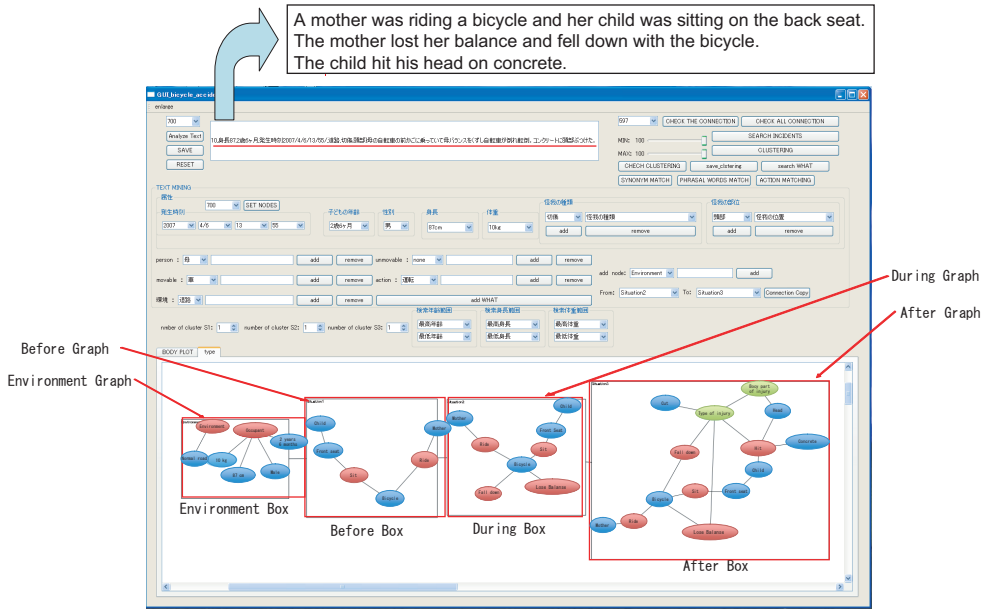


Fig. 4. Example of graph structuralization

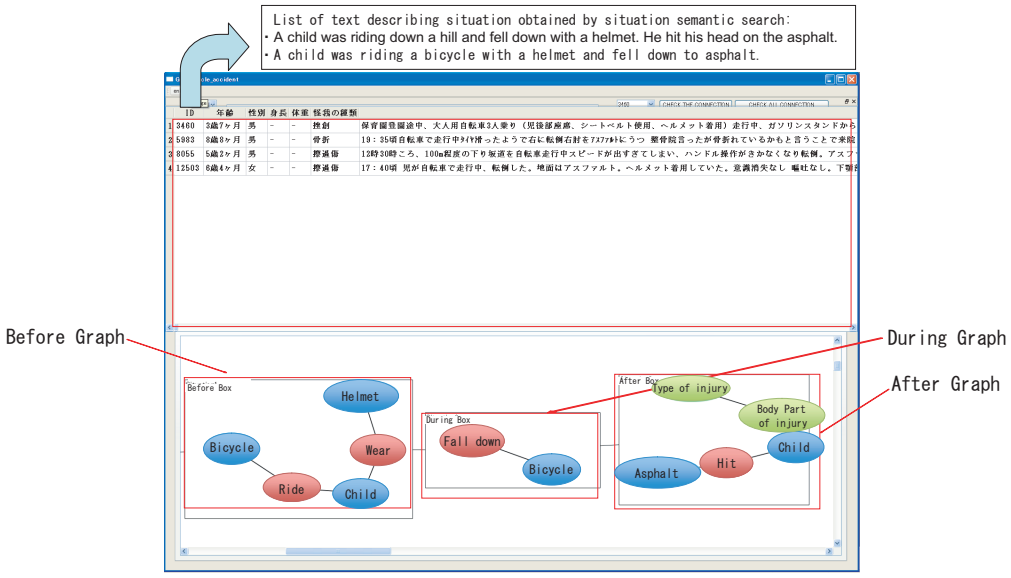


Fig. 5. Example of the result of the situation semantic search

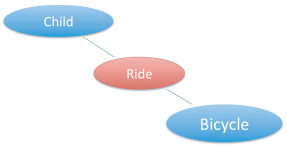


Fig. 2. Example of a Before graph (G_B)

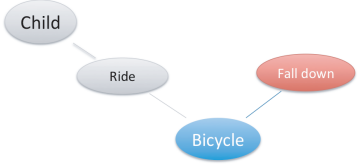


Fig. 3. Example of a unique During graph (G'_D)

phases, namely, the connection between a Before situation and a During situation, the connection among a Before situation,

a During situation, and an After situation, and so forth. In the developed software, the connections are displayed as edges. In addition to this visualization, we use a function that shows the ratio of the number of incidents belonging to the typical situation to the number of all incidents. The visualization allows us to understand which are the components of each typical situation and which components are more important than others.

3 Evaluation of the Situational Analysis System

We performed experiments to evaluate the effectiveness of the developed system by using the real data of 818 child-bicycle incidents, collected at the National Center of Child Health and Development [13].

3.1 Dataset for Evaluation

The 818 child-bicycle incident data include the attributes of occupants, injuries, date of incidents, and site where incidents happened, as well as free-description sentences about the incident situations. The contents in the four situation graphs (Environment, Before, During, and After) are as follows. An Environment graph is created by using the "attributes of an occupant and the environment." A Before graph is created by finding information on the "situation before the incidents" from the free-description sentence. A During graph is created by finding information on the "situation during the incidents" from the free-description sentence. An After graph is created by combining information on the "situation after the incidents" extracted from the free-description sentence, "type of injuries and body parts of injuries" and "which actions caused the incidents."

3.2 Creation of Situation Database

As a specific example of graph structuralization, we show a graph-structuralized situation with one item of data. Figure 4 shows the GUI of the developed software system for the following accident case. Injured child's age: 2 years and 6 months old, Sex: male, Height: 87 cm, Weight: 10 kg, Type of Injury: cut, Body Part of Injury: head, Date and Time of Incident: 13:55 on 6th July 2007, Site of Incident: normal road, Free-description sentence: "A mother was riding a bicycle, and her child was sitting on the back seat. The mother lost her balance and fell down with the bicycle. The child hit his head on concrete." The area at the bottom of Fig. 4 shows the four created graphs (Environment, Before, During, and After) in this case.

3.3 Evaluation of Situation Semantic Search

We conducted an efficacy analysis on the situation semantic search. For the setup, we set Sim_{min}=1. For the search, we graph-structuralized the situation "A child rode a bicycle with a helmet and fell down. After that, he hit his body on concrete" as a condition setting. Figure 5 shows the results at the top and the graph structure for the condition setting at the bottom. We calculated Precision (P), Recall (R), and F-measure (F) in Eqs. (9), (10), and (11).

$$R = \frac{w}{(w+x)}, \tag{9}$$

$$P = \frac{w}{(w+y)}, \tag{10}$$

$$F = \frac{2PR}{(P+R)}, \tag{11}$$

where w is the number of correct documents, x is the number of unexpected documents, and y is the number of missing documents.

The evaluation result is shown in Fig. 6. This figure indicates that by setting Sim_{min}=1 we can search the situation data that matches completely the given situation graph and we can also obtain similar situation data by changing the Sim_{min} value from 0 to1.

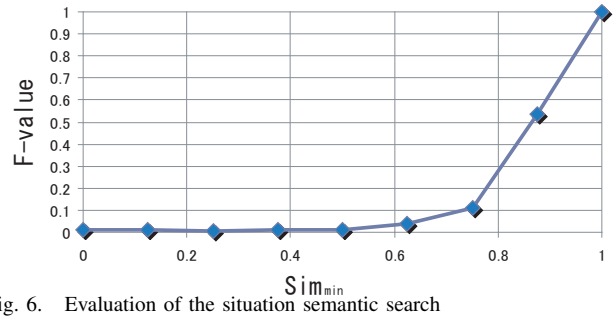


Fig. 6. Evaluation of the situation semantic search

3.4 Effectiveness of Situation Clustering

The developed system supports various kinds of clustering. We show a few examples of clustering by the supported functions. Figure 7 shows an example of the polar dendrogram obtained by hierarchical clustering for the Before graphs. Figures 8 and 9 show the results of clustering using the group average method for the Before graphs and the During graphs, respectively. In the two figures, the number of clusters is 11 for the situational analysis for the Before graphs and 15 for the situational analysis for the During graphs. The number of clusters was determined by the authors by considering the free-description sentence in each cluster. Thus, the system can conduct graph-structuralization based clustering.

We evaluated the effectiveness of situation clustering using the F values. First, we define a typical situation graph for each cluster ID by using the clustering results. Second, we conduct a situation semantic search by giving this typical situation graph to the system as a search query. Then, we can obtain the complete set of situation graphs corresponding to the search query. By comparing this complete set and the results of clustering, we calculate the P, R, and F values. Table I shows the evaluation of clustering for the Before graphs. The F values of the table suggest that the average performance is high in the case of the clustering results shown in Fig. 8. Since this performance depends on the number of clusters that the user gives, the user should set it adequately in actual use; namely, the user should change the grain size of each cluster

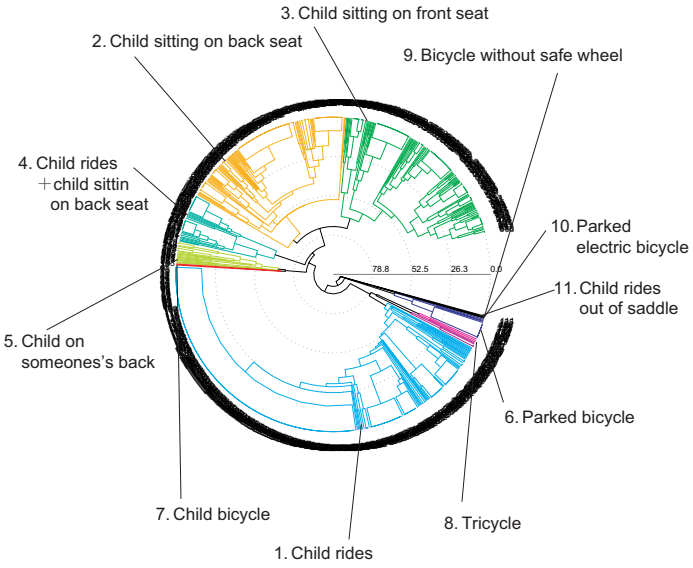


Fig. 7. Polar dendrogram of situation clustering for the Before graphs (number of clusters = 11)

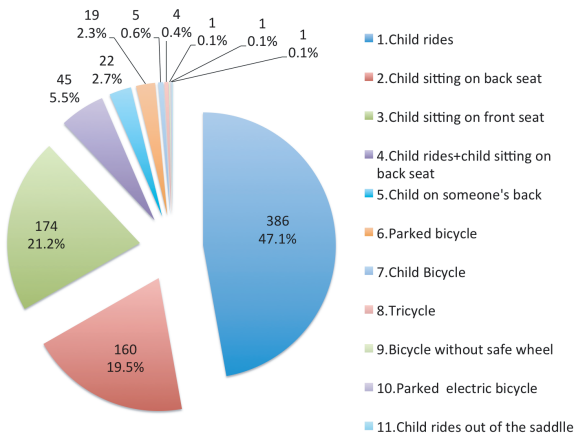


Fig. 8. Result of situation clustering for the Before graphs (number of clusters = 11)

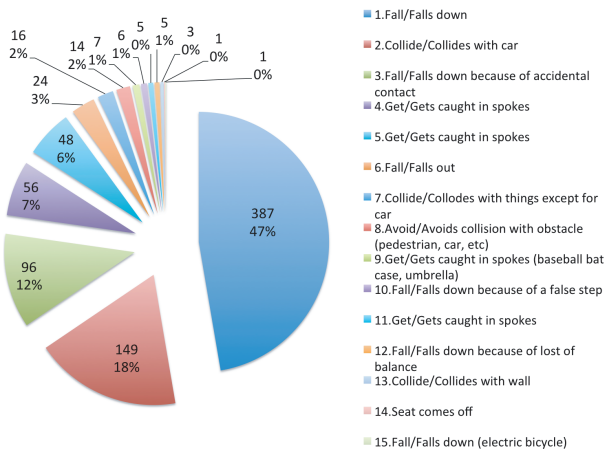


Fig. 9. Result of situation clustering in the case of the During graphs (number of clusters = 15)

so that the user can have an insight into each clustered situation from the viewpoint of injury prevention and situation control.

TABLE I
EVALUATION OF CLUSTERING FOR THE BEFORE GRAPHS

Cluster ID	1	2	3	4	5	6
P (Precision)	0.995	0.994	0.856	0.889	0.273	1.000
R (Recall)	0.948	0.975	0.797	0.068	0.857	0.415
F (F-measure)	0.971	0.985	0.825	0.126	0.414	0.253
Cluster ID	7	8	9	10	11	Average
P (Precision)	1.000	1.000	1.000	1.000	1.000	0.910
R (Recall)	0.833	1.000	1.000	1.000	1.000	0.874
F (F-measure)	0.909	1.000	1.000	1.000	1.000	0.771

3.5 Effectiveness of Visualization of a Typical Situation Process

As a specific example of a typical situation process analysis, we show in Figure 10 the visualization of a typical situation process with the results of clustering we obtained above. The three boxes indicate typical Before situations, During situations, and After situations. The red nodes in each box indicate the most frequent situation. In this figure, for example, we can find the following typical processes. ”Get/Gets caught in spokes (baseball bat case, umbrella)” in a During situation, which is colored yellow in Fig. 10, is connected to two Before situations, ”Child rides” and ”Child sitting on back seat.” The During situation is also connected to the three After situations of ”Bruises,” ”Hit/Hits body” and ”Non-categorized.” Thus, this visualization allows us to understand the components of each typical situation and which components are more important than others.

Using the function of the typical situational analysis, for example, we identified typical situations such as ”a child rides a bicycle and collides with a car,” ”a child drives a bicycle and falls down,” ”someone rides a bicycle with a child on the back seat. The leg of the child gets caught in the spokes,” and ”someone rides a bicycle with a child on the front seat, the bicycle falls down because of accidental contact.”

4 Conclusion

In this research, as a new situational analysis system to extract situation structures from a large number of situation data, we proposed a new situational system that consists of two basic functions: a function for graph structuralization and a function for situation mining based on the situation graph data. The feature of the system lies in formulating a situational analysis as a graph structuralization analysis. We implemented functions for a situation semantic search, linkage analysis of a situation, situation clustering and visualization of a typical situation process. To evaluate the effectiveness of the developed system, we analyzed the real data of 818 child-bicycle incidents. Using this system, we created a database of situation graphs by transforming the 818 bicycle incident data into situation graphs. With the situation semantic search, users can search similar situation data, which is difficult by conventional keyword search methods or text mining methods. Situation clustering is implemented by applying a hierarchical clustering method to the situation graph data. This function

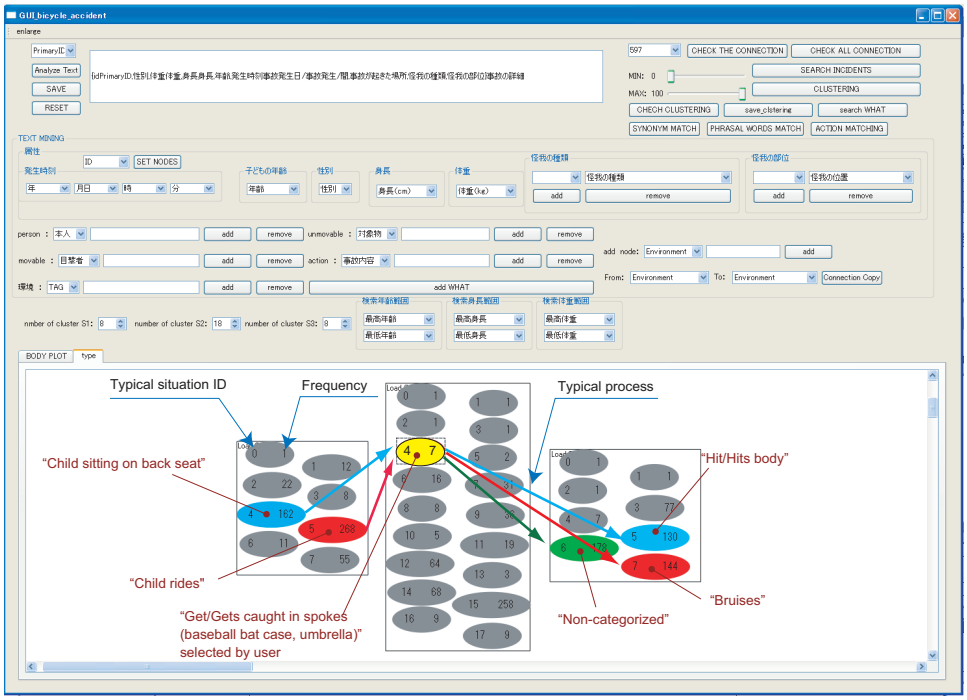


Fig. 10. Visualization of a typical situation process

allows the user to grasp the typical situation of each phase of situations, such as situations before an accident, during an accident, and after an incident. Based on the function of situation clustering, the developed system can visualize typical situation processes.

For future work, we plan to develop a graph-structuralization supporting function. This function will help users to input necessary data for prevention. For example, if a user tries to graph-structuralize an incident and forgets to input information of helmet use status, this new function will let the user know about the necessity of helmet use information. This function will enhance the quality of graph-structuralized situations. In addition, we also plan to make an open database of various incidents. By making a database of graph-structuralized situations of various incidents in accordance with various products or various places, and by sharing the information of situational analyses, we hope to make society safer and more cooperative for injury prevention.

REFERENCES

[1] ISO 3100: 2009 Risk Management-Principles and Guidelines, 2009.
[2] ISO/ICE, Guide 51 Safety Aspects-Guidelines for Their Inclusion in Standards, 1999.
[3] JST Failure Knowledge Database, http://shippai.jst.go.jp/fkd/Search.
[4] Y. Hatamura, Learning from Design Failure, Springer, 2009.
[5] World Health Organization, World Report on Child Injury Prevention, 2008.
[6] J. Shawe-Taylor and N. Christianini, "Kernel Method for Pattern Analysis," Cambridge University Press, 2004.
[7] T. Garther, Kernels for Structured Data, World Scientific Publishing Co. Pte. Ltd., 2008.
[8] P. Papadimitriou, A. Dasdan, and H. Garcia-Molina, "Web Graph Similarity for Anomaly Detection," Journal of Internet Services and Applications, Volume 1 (1). pp. 19-30, 2010.
[9] A. Mislove, M. Marcon, K. P. Gummandi, P. Druschel, and B. Bhattacharjee, "Measurement and Analysis of Online Social Networks," the 5th ACM/USENIX Internet Measurement Conference (IMC'07), San Diego, CA, October 2007.
[10] D. W. Mount, "Sequence and Genome Analysis," Cold Spring Harbor Laboratory Press, 2004.
[11] B. S. Everitt and T. Hothorn, A Handbook of Statistical Analyses Using R, Second Edition, Chapman and Hall/CRC, 2009.
[12] MeCab, http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html.
[13] the National Center of Child Health and Development, http://www.ncchd.go.jp/English/Englishtop.htm.

A new relevance feedback approach for multimedia retrieval

Hanen Karamti[1], Mohamed Tmar[1], and Anis Benammar[2]
[1]MIRACL Laboratory, City ons Sfax, University of Sfax, B.P.3023 Sfax TUNISIA
[2]REGIM Laboratory, ENIS Soukra km 3.5, University of Sfax, B.P.3038 Sfax TUNISIA

Abstract— *Compared to most traditional search engines, information retrieval systems have been devoted in the past few years to relevance feedback (RF). With RF, the user can indicate which documents he finds relevant to his search. RF is an effective solution to improve performance of information retrieval, particularly in multimedia retrieval (image, video).*
In this paper we propose a new approach to relevance feedback in video retrieval. We start out by adapting the standard Rocchio [1], usually used in textual information retrieval, for video retrieval. This adaptation requires the transformation of image retrieval model based on low-level visual features to a vector space model. Next, We propose an automatic query expansion approach used for image retrieval which operates on high-level features. This approach is based on knowledge resources such as the ontology and the conceptual graph.
The proposed approach is evaluated quantitatively and qualitatively using shots from videos collection of the TRECVID10 [2]. The obtained results show the effectiveness of our contributions.

Keywords: Relavance Feedback; Vectorization; Rocchio; Query expansion; Conceptual graph; Ontology

1. Introduction

The main objective of video retrieval is to retrieve all images/videos that are similar to a given query in a database of images/videos. The first content-based image retrieval (CBIR) systems [17] propose automatic retrieval methods based on low-level features (color, texture, shape...). These systems allow the processing of images queries, but they do not make it possible to search for images based on their semantic content. This problem is called the *semantic gap*. A significant improvement of the performance of CBIR systems can be achieved by using relevance feedback, a technique that allows the user to rate the search results.
Since 1970, the relevance feedback mechanisms have been widely deployed in text retrieval [8], [3]. However, their integration into the CBIR systems has been little studied. The CBIR systems based on text retrieval techniques propose solutions based on semantic relations [13], [2] using high-level features. The problem of these systems is due to both ambiguity of the user-supplied query (keywords) and the concepts used to describe images.
In this paper, we adapt the text retrieval relevance feedback techniques to video retrieval. Our first contribution is to adapt the standard *Rocchio*. This adaptation requires applying a *vectorization* method initially used in text retrieval [20]. Our second contribution is to adapt a *query expansion* technique using knowledge resources such as the ontology and the conceptual graph.
This paper is structured as follows: in section 2, we study the various approaches to retrieval with relevance feedback in the multimedia data. In section 3 we describe our approach to relevance feedback. Experiments and results are described in section 4. Finally, section 5 contains discussion and further research directions.

[1]Rocchio standard is based on a method of relevance feedback found in information retrieval systems which stemmed from the SMART Information Retrieval System around the year 1970.

[2]The definitive information about this collection can be found at the NIST TREC Video Track web site: http://www-nlpir.nist.gov/projects/trecvid.

2. Related work

The idea behind relevance feedback is to take the results that are initially returned from a given query and to use information about whether or not those results are relevant to perform a new query.
Relevance feedback has been shown as an effective technique with different retrieval models [2], [7]. In the vector space model, RF is usually carried out using Rocchio algorithm [8], [23], which forms a new query vector(q_{new}) from an inital query (represented by a vector noted q_{old}) by maximizing its similarity to relevant documents (designated by R) and minimizing its similarity to irrelevant documents (designated by S).

$$q_{new} = \alpha q_{old} + \beta \frac{1}{|R|} \sum_{d \in R} d - \gamma \frac{1}{|S|} \sum_{d \in S} d \qquad (1)$$

Where d is document, α, β and γ are positive constants.
The RF method in probabilistic models is to select weighted terms based on Robertson/Sparck-Jones technique (Mars [22]).
In CBIR, the main idea is to propose a mechanism that works with positive and negative images which are returned according to Robertson/Sparck-Jones weight order [19].
Recently, a RF track was used by CBIR through learning models such as the Support Vector Machine (*SVM*). *SVM* is the most popular method; it is used to classify the positive and negative images [16].

In this context, we distinguish the Tow-class SVM, i.e. positive images belong to a class and negative images belong to another class, and both are divided by a hyperplane. We distinguish also the multi-class SVM where the positive and negative images belong to several classes. When they are mixed, it is difficult to find a hyperplane between two classes [4]. This method works with the regions [9], i.e. the user can select the positive and negative regions in an image. After applying relevance feedback, the similarity of each region changes, and the similarity of the image is the combination of all regions.
Further work is based on the negative relevance feedback through learning negative examples [18].
Other researches are based on descriptors [12]. The role of those approaches is to determine the importance of each descriptor involved in the CBIR. In the beginning, the importance of each descriptor is initialized with the same value. Then it is updated throughout the retrieval session.
During each iteration, the user evaluates the displayed images and decides whether they are relevant or irrelevant to his need. The system uses the evaluated images as training samples to successively refine the learning model and give a better retrieval quality in the subsequent iteration. The problem of this process is that it requires an important number of iterations to achieve reasonable performance which can be time-consuming and tedious. All these aspects urge us to use an automatic approach of relevance feedback without involving the user judgment, such method is called '*pseudo relevance feedback*' [21]. In this approach, the relevant documents are identified automatically by assuming that the top-ranked documents are relevant [18], [10].
To improve the search quality, another promising direction is to introduce the query expansion based on concepts related to the query [1], [6] to assist researchers to refine their search process.

3. A new approach of relevance feedback

We work on the keyframes of video shots, we assume that if the keyframe is relevant then the shot containing this image is also relevant.

3.1 Adaptation of the standard Rocchio

To adapt the standard Rocchio, we have to transform the matching images model to a vector space model. This transformation is called 'vectorization'. The idea is to convert the space of low-level features extracted from images to a vectors.

3.1.1 Vector space model

In information retrieval, the vector space model represents the document and the queries on a vector space with dimension n. The $\vec{d} = (w(1,d), w(2,d) \cdots w(n,d))$ is the vector associated to the document d, when, $w(i,d) \in [0,1]$ is the weight often t_i in the document d.
It is the same for the query, q represents the vector noted $\vec{q} = (w(1,q), w(2,q) \cdots w(n,q))$. When $w(i,q) \in [0,1]$ is the weight often t_i in query q. The term weighting is defined with the term frequency (Tf) and the inverse document frequency (Idf) [8].
The most popular similarity measures are:
Cosinus:

$$Cos(q,d) = \frac{\sum_{i=1}^{n} w(i,q) \times w(i,d)}{\sqrt{\sum_{i=1}^{n} w(i,q)^2} \times \sqrt{\sum_{i=1}^{n} w(i,d)^2}} \quad (2)$$

Jaccard:

$$C(q,d) = \frac{\sum_{i=1}^{n} w(i,q) \times w(i,d)}{\sqrt{\sum_{i=1}^{n} w(i,q)^2} + \sqrt{\sum_{i=1}^{n} w(i,d)^2} - \sum_{i=1}^{n} w(i,q) \times w(i,d)} \quad (3)$$

Dice:

$$Dice(q,d) = \frac{2 \times \sum_{i=1}^{n} w(i,q) \times w(i,d)}{\sqrt{\sum_{i=1}^{n} w(i,q)^2} + \sqrt{\sum_{i=1}^{n} w(i,d)^2}} \quad (4)$$

OverLap:

$$Overlap(q,d) = \frac{\sum_{i=1}^{n} w(i,q) \times w(i,d)}{min\left(\sqrt{\sum_{i=1}^{n} w(i,q)^2}, \sqrt{\sum_{i=1}^{n} w(i,d)^2}\right)} \quad (5)$$

3.1.2 Vectorization process

As the classical vector space model, we can represent the image query Q by a vector of low-leval features:

$$\vec{Q} = (q_{c1}, q_{c2} \ldots q_{ck}, q_{f1}, q_{f2} \ldots q_{fp})$$

Where k is the number of features extracted by the descriptor *CLD* [3] (Color Layout Descriptor), p is the number of features extracted by the descriptor *EHD* [4] (Edge Histogram Descriptor), q_c is the color feature extracted by CLD descriptor and q_f is the shape feature extracted by EHD descriptor.

[3] A color layout descriptor (CLD) is designed to capture the spatial distribution of color in an image. The feature extraction process consists of two parts; grid based representative color selection and discrete cosine transform with quantization [11].

[4] Edge Histogram Descriptor (EHD) is proposed for MPEG-7 expresses only the local edge distribution in the image [5].

The keyframes in the database are represented by a $n * m$ matrix M:

$$M = \begin{pmatrix} I_{(1,c1)} & ... & I_{(1,ck)} & I_{(1,f1)} & ... & I_{(1,fp)} \\ & & & . & & \\ & & & . & & \\ & & & . & & \\ I_{(n,c1)} & ... & I_{(n,ck)} & I_{(n,f1)} & ... & I_{(n,fp)} \end{pmatrix}$$

$$= \begin{pmatrix} I_{1,1} & I_{1,2} & I_{1,3} & ... & I_{1,m} \\ & & . & & \\ & & . & & \\ & & . & & \\ I_{n,1} & I_{n,2} & I_{n,3} & ... & I_{n,m} \end{pmatrix}$$

Where n is the number of database keyframes, $m = k + p$ is the number of matrix columns, $I(j, c_k)$ is the k^{th} color feature c extracted from keyframe j and $I(j, f_k)$ is the k^{th} shape feature f extracted from keyframe j with $j \in \{1, 2...n\}$.
The initial i keyframes returned by the system are represented by the reference $i * m$ matrix R:

$$R = \begin{pmatrix} I_{1,1} & I_{1,2} & I_{1,3} & ... & I_{1,m} \\ & & . & & \\ & & . & & \\ & & . & & \\ I_{i,1} & I_{i,2} & I_{i,3} & ... & I_{i,m} \end{pmatrix}$$

To calculate the similarity between query and keyframes, we must, on the one hand, calculate the similarity between query and reference keyframes, with $S(Q, I_j)$ the score of similarity between query Q and reference keyframe I_j where $j \in \{1, 2...i\}$:

$$Sim(\vec{Q}, R) = (S(Q, I_1), S(Q, I_2)...S(Q, I_i))$$

On the other hand, we must calculate the similarity between reference keyframes and database keyframes, where $S(I_i, I_j)$ is the similarity between the i^{th} image in the database and the j^{th} reference keyframe:

$$Sim(M, R) = \begin{pmatrix} S(I_1, I_1) & S(I_1, I_2) & ... & S(I_1, I_i) \\ & & . & \\ & & . & \\ & & . & \\ S(I_n, I_1) & S(I_n, I_2) & ... & S(I_n, I_i) \end{pmatrix}$$

Then, we can construct the new query vector $\vec{Q_{new}}$ defined by the vector $\vec{Sim(Q, R)}$ in the new search space $Sim(M, R)$ containing the new keyframes ($\vec{I_{new}}1, \vec{I_{new}}2...\vec{I_{new}}n$) with the same size as $\vec{Q_{new}}$. Therefore, we have:

$$\vec{I_{new}}n = (S(I_n, I_1), S(I_n, I_2), ..., S(I_n, I_i)).$$

So a new search is triggered and a new vector result $\vec{R_2}$ is constructed. To calculate the distance between vectors, we use the Overlap (equation 5) metric:

$$\vec{R_2} = (Overlap(\vec{Q_{new}}, \vec{I_{new}}1)...Overlap(\vec{Q_{new}}, \vec{I_{new}}n)) \quad (6)$$

3.1.3 Integration of standard Rocchio

To improve the results returned by vectorization technique, we integrate the standard Rocchio (equation1). We put γ to 0, and we evaluates the system with different values of α and β. The must appropriate values of α and β are respectively 0.3 and 0.7.
The first n images in the vector result $\vec{R_2}$ will be added to the initial query to build a new query (equation 6).
This process will be reiterated until the user satisfaction.

3.2 Conceptual query expansion for image search

In this section, we are going to explore how semantic relations can be used to expand a query for concepts describing an image [14]. Query expansion consists in adding some synonym or relative words into the query set of original keywords to improve the recall and the precision of information retrieval. Traditional methods of query expansion did not make adequate use of semantic relations between query keywords. They often give bad results on recall and precision. These methods can be improved by using knowledge resources such as ontology and conceptual graph.
In this paper, a novel approach for query expansion is presented. The main idea of the approach is to add to the query a set of keywords which are extracted from the representation of the relations between concepts. In the first step, a conceptual graph is constructed based on LSCOM [5] ontology.
In the second step, keywords are extracted by the intersection of connected subgraphs representing each video from the result. Finally, all keywords extracted will be added to the initial query. The features extracted from images are saved in indexes. We also used an XML file containing a set of concepts derived from semantic indexing key frames [15].

3.2.1 Construction of conceptual graph

A Conceptual graph is a graph where nodes are concepts and edges indicate the relationship between them. In this section, we explain the steps we followed to create this graph. We have constructed the conceptual graph using a lexicon of 130 concepts from the LSCOM ontology, used during the last session of the evaluation campaign TRECVID10. The table 3.1 shows an overview of LSCOM concepts.
Figure 1 shows an excerpt of the LSCOM ontology representing the various semantic relations between some concepts of TRECVID10. From these relations, the conceptual graph defined by Figure 2 is constructed.

[5] http://www.lscom.org/

Table 1: Excerpt of concepts used in TRECVID2010

TV10_ID	LSCOM_ID	LSCOM_Name	Definition
001	149	Actor	One or more television or movie actors or actresses
002	181	Adult	Shots showing a person over the age of 18
003	218	Airplane	Shots of an airplane
004	125	Airplane_Flying	An airplane flying in the sky
005	1062	Anchorperson	Anchorperson
006	202	Animal	Shots depicting an animal (no humans)
007	246	Asian_People	People of Asian ethnicity

Bicycles implies Ground_Vehicles
Bicycling implies Bicycles
Birds implies Animal
Basketball implies Person
Basketball implies Sports
Cats implies Animal
Cows implies Animal
Cheering implies Person
Conference_Room implies Indoor

Fig. 1: Excerpt of semantic relations resulting from LSCOM ontology

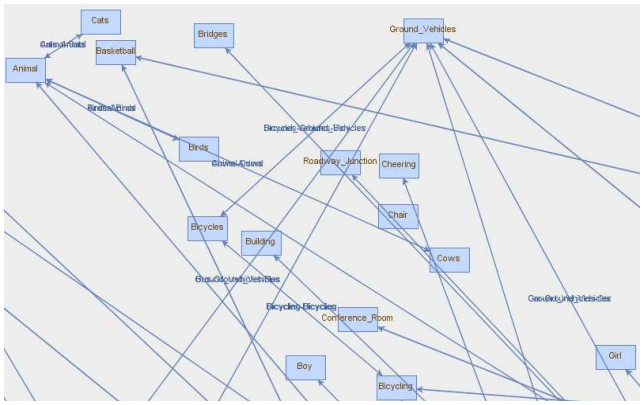


Fig. 2: Conceptual graph

3.2.2 Integration of conceptual graphs for query expansion

After construction of the conceptual graph, we extract the connected subgraphs among all concepts of each video shot returned in the result. In fact, we search a path relating all the concepts of shot video. The relationships between concepts bring new concepts which are not defined in the shot. Furthermore, the connected subgraphs should be intersecting. The intersection between them is a set of new concepts which will be added to the initial query.
Let $G(N, R)$ the conceptual graph where $N = \{c_1, c_2...c_n\}$ is the set of concepts and $R \subset N^2$ is the set of edges. The video shot returned by the initial result is represented by V, where $V \subset N$.
We say that a shot V is a connected subgraph if
$\forall c_i, c_j \in V \exists k \in \mathbb{N}, c'_1, c'_2...c'_k \in N$ and $(c_i, c'_1) \in R$ and $\forall j \in \{1, 2...k-1\}, (c_j, c_{j+1} \in R)$ and $(c'_k, c_j) \in R$.
Let $Comp_1$, $Comp_2$ ... $Comp_n$ be the connected subgraphs of G representing returned video shots in the result. $Comp_i$ is the connected subgraph linking all the concepts of video shot i , i.e. the concepts from i and the intermediary concepts between them. The query Q is represented by a subgraph in G where $Q = \{q_1, q_2...q_n\}$ is the set of query concepts.
If Q is a connected subgraph in G then the new query T_{new} is defined by:

$$T_{new} = Comp_1 \cap Comp_2 \cap ... \cap Comp_n \cap Q$$
$$= \left(\bigcap_{i=1}^{n} Comp_i \right) \cap Q$$

If Q is not a connected subgraph in G then:

$$T_{new} = \bigcap_{i=1}^{n} Comp_i$$

4. Experiments

We evaluate our relevance feedback approach into TRECVID10 collection. We limit our experiments to 186 video designated by 6915 keyframes.
The table 2 defines the queries used in our tests, theirs precision values and recall values returned after an initial search.

Table 2: Recall values and precision values returned by an initial research with the queries defined

Queries	Relevant images	Recall	Precision
"Male_Person" 	117	0.22	0.41
"Building" 	65	0.32	0.3
"Car" 	30	0.76	0.35

Table 3: Evaluation of the recall and precision according to the number of reference images

n	"*Male_Person*"		"*Building*"		"*Car*"	
	Recall	Precision	Recall	Precision	Recall	Precision
10	0.29	0.43	0.35	0.36	0.73	0.34
15	0.27	0.38	0.37	0.38	0.80	0.38
20	0.28	0.40	0.43	0.43	0.77	0.43
25	0.30	0.44	0.48	0.48	0.77	0.36
30	0.28	0.41	0.48	0.48	0.77	0.36
35	0.31	0.44	0.50	0.50	0.83	0.40
40	0.31	0.44	0.55	0.56	0.83	0.40
45	0.31	0.44	0.55	0.56	0.87	0.40

4.1 Assessment of vectorization approach and Rocchio adaptation

4.1.1 choice of references number

Table 3 shows that the recall increases proportionally to the number of references (n) for the "*Building*" query, but it depends on n for "*Male_Person*" and "*Car*" queries. We show that the recall decreases when n varies between 10 and 30. Thus, it takes a growing path. Therefore, we can conclude that the recall depends on the chosen number of references.
To identify the most appropriate value of n, we calculate the precision of these three queries when n is between 10 and 45. Table 3 shows that when n reaches the value 40, the precision becomes constant.
In what follows, we set n to 40.

4.1.2 Evaluation Metrics

To choose the best similarity measure, we evaluate four arithmetic functions: Cosinus, Jaccard, Dice and Overlap. We notice a large number of images returned from all the queries using the Overlap function (figure 3). Therefore, we can conclude that the system returns the most relevant images using the Overlap function to calculate the similarity between the query and images vector.

4.1.3 Assessment Rocchio

For the Rocchio adaptation, we evaluated the α and β values for both queries "*Male_Person*" and "*Building*". In Table 4, we find that "*Building*" query has the best precision results (0.67) where α and β values equal respectively 0.3 and 0.7. Another precision value is 0.65 when α equals 0.7 and β equals 0.3. For "*Male_person*" query, the high value of precision equals 0.42 when α and β are respectively 0.8 and 0.2 or 0.7 and 0.3. From these values, we can conclude that when α is set to 0.7 and β to 0.3, the precision value is better than the initial result search (precision $= 0.41$ for *Male_Person*" query, precision $= 0.3$ for "*Building*" query).

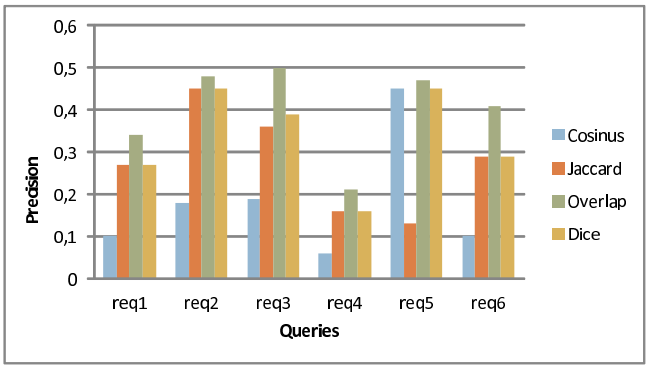


Fig. 3: Evaluation of the similarity metrics

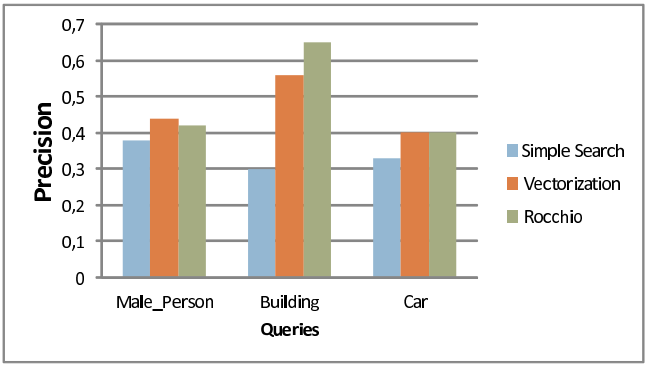


Fig. 4: Evaluation of the Rocchio technique compared to others research techniques

The recall also reaches its maximum value for the same values of α and β.
Figure 4 summarizes the quality of the system responses before and after the RF (by integrating the vectorization technique only or by integrating the Rocchio standard with vectorization technique). From this histogram, we notice that when we apply the standard Rocchio, the precision value is improved comparised with the initial result returned by the system, but it remains dependent of the quality of

Table 4: Variation of values of the α and β for the *Rocchio* adaptation

"Building"				"Male_Person"			
α	β	Recall	Precision	α	β	Recall	Precision
0.5	0.5	0.62	0.58	0.5	0.5	0.23	0.34
0.4	0.6	0.51	0.49	0.4	0.6	0.27	0.35
0.3	0.7	0.68	0.67	0.3	0.7	0.19	0.27
0.2	0.8	0.54	0.49	0.2	0.8	0.19	0.27
0.1	0.9	0.62	0.59	0.1	0.9	0.19	0.28
0.9	0.1	0.54	0.51	0.9	0.1	0.2	0.3
0.8	0.2	0.51	0.48	0.8	0.2	0.29	0.42
0.7	0.3	0.68	0.65	0.7	0.3	0.29	0.42
0.6	0.4	0.26	0.51	0.6	0.4	0.19	0.27

Table 5: Evaluation of the system retrieval after integration of semantic relations

Queries	Recall(RI)	Precision(RI)	Recall(RF)	Precision(RF)
Male_Person	0.22	0.41	0.34	0.49
Building	0.32	0.3	0.32	0.41
Person in the car	0.87	0.27	0.87	0.36
Person watch $news\_studio$	0.62	0.5	0.9	0.58

[a] RI: Initial search, RF: search with RF

vectorization results.

4.2 Assessment of expansion query

The quality of the result after the integration of semantic relations is evaluated by two queries. We see that recall value and precision value increased compared with the initial result, which proves the effectiness of our approach (see table 5).
Table 6 shows the new concepts found by query expansion technique.

Queries	Relevant terms
Person	Adult, Single_Person
Car	Vehicle, Ground_Vehicle
Bicycling	Bicycles

Table 6: Relevant terms added to the query

5. Conclusion

In this article, we presented a study of the context of relevance feedback for multimedia retrieval. The mechanisms deriving from RF have been widely used in textual information retrieval. However, their induction in the process of video retrieval is underdeveloped. In addition, video data causes famous problem of *semantic gap.*
Our approach is to adapt a standard Rocchio, usually used if textual information retrieval, for the multimedia information retrieval. This adaptation required additional phases such as vectorization. We improve the semantic interpretation of queries through the induction of such knowledge structures such as ontologie and conceptual graph. Such structures were used to expand the user query to improve its expression.
To increase the effectiveness of our approach, it is possible to make some improvements and changes. It would be interesting to explore the LSI concept in the multimedia retrieval process. As for the knowledge structures, in order to induce the process of query expansion, it would be particularly interesting to take into account the context notion that can be deduced from existing correlations between concepts.

References

[1] Dr. A.K.Sharma and P.Gulati. Article: Ontology driven query expansion for better image retrieval. *International Journal of Computer Applications*, 5(10):33–37, August 2010. Published By Foundation of Computer Science.

[2] D.Heesch. A survey of browsing models for content based image retrieval. *Multimedia Tools Appl.*, 40(2):261–284, November 2008.

[3] D.Hiemstra and S.E.Robertson. Relevance feedback for best match term weighting algorithms in information retrieval. In A. Smeaton and J. Callan, editors, *Proceedings of the Joint DELOS-NSF Workshop on Personalisation and Recommender Systems in Digital Libraries*, volume 01 of *ERCIM Workshop Proceedings*, pages 37–42. European Research Consortium for Informatics and Mathematics, June 2001.

[4] D.H.Kim, J.W.Song, J.H.Lee, and B.GH.Choi. Support vector machine learning for region-based image retrieval with relevance feedback. *ETRI Journal*, 29(5):700 – 702, 2007.

[5] D.K.Park, Y.S.Jeon, and C.S.Won. Efficient use of local edge histogram descriptor. In *Proceedings of the 2000 ACM workshops on Multimedia*, MULTIMEDIA '00, pages 51–54, New York, NY, USA, 2000. ACM.

[6] D.Milne, I.H.Witten, and D.M.Nichols. A knowledge-based search engine powered by wikipedia. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 445–454, New York, NY, USA, 2007. ACM.

[7] I.Kamoun Fourati, M.Tmar, and A.Ben Hamadou. Structural relevance feedback in xml retrieval. In *FQAS'09*, pages 168–178, 2009.

[8] G.Salton. *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.

[9] J.Luo and A.Nascimento. Content-based sub-image retrieval using relevance feedback. In *Proceedings of the 2nd ACM international workshop on Multimedia databases*, MMDB '04, pages 2–9, New York, NY, USA, 2004. ACM.

[10] K.S.Lee, W.B.Croft, and J.Allan. A cluster-based resampling method for pseudo-relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 235–242, New York, NY, USA, 2008. ACM.

[11] L.Cieplinski. Mpeg-7 color descriptors and their applications. In *Proceedings of the 9th International Conference on Computer Analysis of Images and Patterns*, CAIP '01, pages 11–20, London, UK, 2001. Springer-Verlag.

[12] Thi-Lan LE. *Indexation et Recherche de Video pour la Videosurveillance*. PhD thesis, Universite de Nice-Sophia Antipolis, 2009.

[13] M.L.Kherfi, D.Ziou, and A.Bernardi. Image retrieval from the world wide web: Issues, techniques, and systems. *ACM Comput. Surv.*, 36(1):35–67, March 2004.

[14] M.Naphade, R.J.Smith, J.Tesic, SH.Chang, W.Hsu, L.Kennedy, A.Hauptmann, and J.Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91, July 2006.

[15] N.Elleuch, M.Zarka, I.Feki, A.Ben Ammar, and A.Alimi. Regimvid at trecvid 2010: Semantic indexing, 2010.

[16] P.Hong, Q.Tian, and T.Huang. Incorporate support vector machines to content-based image retrieval with relevant feedback. pages 750–753, 2000.

[17] R.Lempel and A.Soffer. Picashow: Pictorial authority search by hyperlinks on the web, 2001.

[18] R.Yan, A.G.Hauptmann, and R.Jin. Negative pseudo-relevance feedback in content-based video retrieval. In *In Proceedings of ACM Multimedia (MM2003*, pages 343–346, 2003.

[19] S.E.Robertson and K.S.Jones. Simple, proven approaches to text retrieval. *Update*, 1997.

[20] V.Claveau, R.Tavenard, and L.Amsaleg. Vectorisation des processus d'appariement document-requête. In *CORIA*, pages 313–324. Centre de Publication Universitaire, 2010.

[21] Y.Lv, CH.Zhai, and W.Chen. A boosting approach to improving pseudo-relevance feedback. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 165–174, New York, NY, USA, 2011. ACM.

[22] Y.Rui, T.S.Huang, and .Mehrotra. Relevance feedback techniques in interactive content-based image retrieval. In *IN STORAGE AND RETRIEVAL FOR IMAGE AND VIDEO DATABASES (SPIE*, pages 25–36, 1998.

[23] Y.Rui, T.S.Huang, M.Ortega, and S.Mehrotra. Relevance feedback: A power tool for interactive content-based image retrieval, 1998.

Dynamic parameter learning method for information retrieval using genetic algorithm

Liping Yang[1], Qinhong Sun[2], Fenxiang Liu[2]
[1,2]College of Computer Science and Engineering, Sanjiang University, Nanjing 210012, China

Abstract *- Parameters setting in the information retrieval systems greatly affects retrieval performance. Those parameters are always data-dependent and sensitive, which may cause the fallibility of experiential values. What is more, due to the lacking of relevant information while retrieving, supervised parameter learning approaches are not applicable. Therefore, there is a great need for an automatic unsupervised parameter learning mechanism. In this paper, we first studied the effectiveness of traditional manual parameter setting with fixed experiential values, which reveals that the previous methods are not feasible or reliable to use widely in practice; then, we proposed a dynamic parameter learning method with genetic algorithm. Experiments have been done on large-scale data sets of TREC7, TREC8 and TREC9 web track collections, each of which reaches 10GB. Results show that the system approaches the best retrieval performance using our dynamic parameter learning.*

Keywords: information retrieval; data-dependent; unsupervised parameter learning; dynamic parameter learning

1 Introduction

The goal of information retrieval systems is predict the correlation between each document and user query. There are different retrieval models with various similarity calculation methods, all of which need to set up a set of parameters. Different parameter values will bring a completely different search results, which plays a decisive impact on system performance. It is an important but unsolved problem to set those parameters.

Recently, most researchers fixed those system parameters with their experience, which are from two ways: 1) According to the principle of similarity calculation, we can get a possible range of parameters. However, this range is usually large, which is difficult to apply to real application. 2) Optimization of system parameters on the training set used directly for the test set, but the data retrieval system parameters and sensitivity, this approach is not entirely reliable, resulting in the instability of the retrieval performance. In addition, due to lack of information retrieval, and thus cannot be guided learning.

In this paper, according to the study of parameter learning, we proposed a non-guided retrieval system parameter dynamic learning. In section 2, According to the experiments on large-scale test data set, we studied the generalization ability of traditional experience the value of setting parameters. In section 3, we proposed credibility assume and a new fitness function based which, we given a retrieval system using dynamic parameter learning with a genetic algorithm. In section 4, we applied the proposed method to the real time data and analyzed the results. The conclusion and our future work is discussed in last section.

2 Generalization of traditional methods

2.1 Fundamental research and data

The probabilistic retrieval model [1] has been proposed since the 1980s, it has been proved to be one of the most effective retrieval models. The model-based experimental system also performs well in well-known standard evaluation of text information retrieval TREC. The basic idea of the probability model to estimate the similarity of a document and a given user query topics, and thus return to the related document [2]. In the classical probabilistic model, the similarity of document and query can be calculated with the formula BM25 [3] and its various variants. One of most famous and widely adopted formula BM25 is shown as

follow:

$$score(D,Q)=\sum_{i=1}^{n} IDF(q_i)\cdot\frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot(1-b+b\cdot\frac{|D|}{avgdl})},$$

$$IDF(q_i)=\log\frac{N-n(q_i)+0.5}{n(q_i)+0.5} \quad (1)$$

In which, $score(D,Q)$ is what we want to calculate the score, that is, to the search the contents of Q in the given document D, with higher scores indicating the degree of the higher. q is a morpheme in a given search content Q. $f(q_i,D)$ is the frequency of a morpheme q_i in a given document D. $|D|$ is the length of a given document D. $avgdl$ is the length of all documents in index. The other two parameters K1 and b are used to adjust the accuracy, under normal circumstances, we take k_1 = 2, b = 0.75. N is the total number of documents in the index. $n(q_i)$ is the total number of the index contains the documentation of the morpheme q_i .This article studies the probabilistic retrieval model, for example, on the basis of BM25 similarity computation formula and experimental verification.

Experiments have been done on large scale data sets of TREC7, TREC8 and TREC9 web track collections [4], each of which reaches 10GB. TREC is a text field of information retrieval of the most important criteria evaluation meeting; it is famous for its large-scale data set of test, evaluation of impartiality and openness. Therefore, the experiments conducted on the data set are highly reliable and convincing.

TREC7 uses WT10g data, which is downloaded from the Internet, about 1.69 million web pages. TREC8 and TREC9 use .GOV data, which includes 1.25 million web pages. Contains HTML and other types of documents such as PDF, TXT, MSword unstructured collection of WT10g and GOV. Their proportion in line with the proportion of the distribution of documents on the Internet can be regarded as the true reflection of life document usage.

TREC7, TREC8 and TREC9 have 50 user queries and a group corresponding to the document collection as evaluation of the standard answer respectively. The TREC organizers of the NIST laboratory assistant Pooling technical [5] manual marked out use it. The two data sets in the document type, document size and content are very different, completely different set of test data.

2.2 Generalization capabilities

2.2.1 Different datasets

Figure 1 lists the retrieved using probabilistic retrieval model BM25 formula in TREC7, and TREC8 data set taken in b for different values of average precision, in which k_1 =2. For the dataset TREC7, MAP is greater than 0.21 when b is from 0.45 to 0.70; while for the dataset TREC8, MAP reaches its peak when b is 0.80.

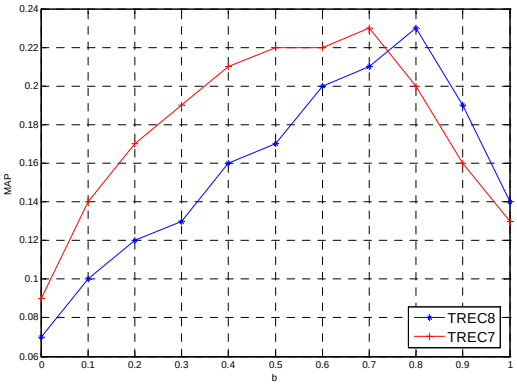


Fig. 1 Retrieval effects by various values of b in different datasets (k_1 =2)

Table 1 shows the parameters to achieve optimal performance on TREC7 or TREC8 experience, to retrieve the performance impact on a data set. The best parameters to use TREC7 experience, the after replaced TREC8 data retrieval MAP and compared to the optimal performance loss of 17.1%; can also cause performance degradation of 7.9%.

Table 1 Effects of using empirical values of b in different datasets (k_1 =2)

Empirical values of b	TREC7		TREC8	
	MAP	Compared with best performance	MAP	Compared with best performance
0.5	0.2105	—	0.1745	-17.1%

0.7	0.1929	-7.9%	0.2094	—

2.2.2 Same dataset with different user queries

Figure 2 shows the experimental results of TREC8 and TREC9. GOV test set. These two data sets are used .GOV dataset, but 50 different user queries. System performance evaluation using the average accuracy of 10 documents avg (10). In the TREC8 data sets, when the b value is 0.2-0.5, the system performance is excellent, especially in the b = 0.4, reaching the maximum value of 0.2517. In TREC9 test set, only when b = 0.6, retrieval performance in order to achieve optimal 0.2765. Similarly, Table 2 gives a set of optimal parameters as the role of experience in the performance of another set of user queries. Compared with the optimal parameters, the use of empirical values were more than 7.6% and 9.1% performance loss.

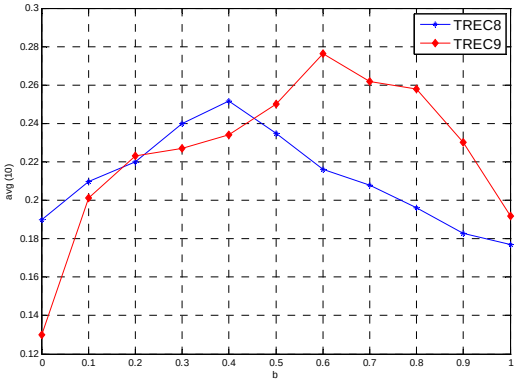


Fig. 2 Retrieval effects by various values of b in the same dataset (k_1 =2)

Table 2 Effects of using empirical values of b in the same dataset (k_1 =2)

Empirical values of b	TREC8		TREC9	
	avg (10)	Compared with best performance	avg (10)	Compared with best performance
0.4	0.2465	—	0.2278	-7.6%
0.6	0.2540	-9.1%	0.2794	—

2.2.3 Different evaluation criteria

Figure 3 shows the value in the TREC9 on the test set, b, respectively, MAP and The avg (10) to evaluate the resulting performance. Avg (10) of the system reaches its maximum 0.2105 when b = 0.5, while the MAP reaches maximum 0.2765when b=0.6. Table 3 shows that the optimal parameters as one of the evaluation criteria used in the TREC9 data experience, another evaluation criteria, system performance will also result in losses.

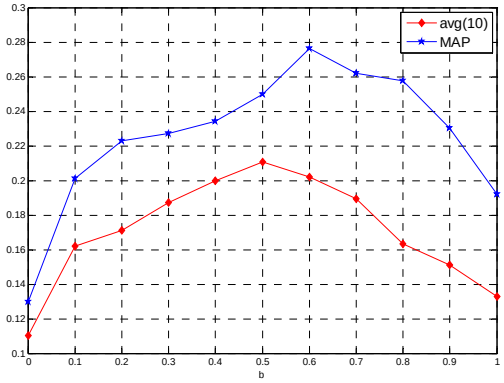


Fig. 3 Retrieval effects by various values of b in the same dataset and user queries with different evaluation criteria (k_1 =2)

Table 3 Effects of using empirical values of b in the same dataset and user queries with different evaluation criteria (k_1 =2)

Empirical values of b	Avg(10)		MAP	
	Performance	Compared with best performance	Performance	Compared with best performance
0.5	0.2105	—	0.2354	-14.9%
0.6	0.1924	-8.6%	0.2765	—

We can see by inspection of the above-mentioned three aspects, have traditionally relied on experience to set the system parameters is not feasible. : 1) for the different data collection be retrieved with the same retrieval model and system for optimum performance parameters may be completely different; 2) even if the same data collection, and the user query, retrieve best the parameter values of the performance may be completely different; 3) use the same data set with the same user query and retrieval performance evaluation criteria to obtain the optimal parameters are also

different.

3 Unsupervised learning methods

Existing retrieval model, using the experience to set the model parameters is not a viable solution! Different values of the parameters greatly affect the retrieval performance of the system on the other hand, during the search, each user query the relevant set of documents is unknown, so a variety of supervised learning method is not available It appears that we need to find an automatic parameter learning method, the query itself from the data and user access to certain auxiliary information to guide the parameter learning.

3.1 Fundamental research

In the field of machine learning, solving a complex problem can be seen to solve the solution is to find the optimal exhaustive of all possible options to compare to find the optimal solution method for solving problems in a small space, you can use! But in solving space, you must use a special artificial intelligence search algorithms, genetic algorithm [6] is an alternative method for information retrieval system parameter adjustment, but also such a large solution space to find the optimal solution of the problem here the optimal solution is the ability to make the system performance to achieve the optimal parameters. Genetic algorithm does not require domain knowledge, adaptability, in previous studies showed good learning outcomes, and thus was chosen as a parameter in these study-learning methods.

Genetic algorithm was first Holland [7] suggested that it is an evolutionary learning method to achieve through an iterative search technique. With the traditional optimization process [8], the genetic algorithm from multiple initial solutions through evolution, while reaching an approximate optimal solution. This process is too simulating the natural biological selection and reproduction of random information exchange to complete. The problem to be solved by the chromosome encoded by multiple genes to describe each chromosome has an adaptation. The value of the fitness function evaluation to adapt to the good or bad and the survival probability function used to determine the chromosome, the higher the fitness, the chromosome is more likely to be selected for breeding.

Thus, the focus of genetic algorithm parameters learning is unknown document collection under the premise of how to define a suitable fitness function for future generations of breeding selection and optimization. It has been studied the use of genetic algorithms in information retrieval [9], but mainly for the relevance feedback techniques rather than the system parameter adjustment. Vrajitoru [10] in TREC5 have used genetic algorithm adjustable parameters, then use the standard answer as to the fitness function. Although this experiment is to get good results, but because the standard answer that "a collection of documents can not be in the actual retrieval, and therefore can not be used in practice. Therefore, using a genetic algorithm for unsupervised parameter learning for information retrieval systems is the first time in this article have been proposed and considered.

Easy to know if you know the standard answer, and the final retrieval performance evaluation criteria, you get the best performance of the brute-force method is parameter learning method for the maximum performance of this study will use this upper limit of performance comparative study.

3.2 New fitness function

Firstly, the credibility of the assumptions of the system model in the search and retrieval of the similarity of the results distribution assumption:

1) The credibility of the assumptions of the retrieval model. Retrieval model to calculate and measure the similarity between document and query results within the scope of the sort the front is basically credible, and credibility is enhanced with the increase of similarity. That is to say, if the similarity of the document retrieval model is very high, then the search results more credible, the retrieval performance of the system is relatively better; sorting after the retrieval system returns relevant documents may be related to the actual conditions do not meet, and even has nothing to do with the user's query.

2) The distinction between assumptions of the search results. Retrieval model, the stronger the similarity distinction of relevant documents, search results is more credible. That is to say, if a retrieval system to get search results relevant documents high similarity, low similarity values without documentation, then the retrieval results are reliable. On the contrary, system similarity values are

relatively close, the system can not well distinguish between relevant and irrelevant information, and thus results not credible.

In fact, the establishment of these two assumptions must meet two basic premises: 1) retrieval model to solve the retrieval problem to a certain extent, but there is no obvious fundamental error! Otherwise, a wrong model or retrieval cents model to get all the results are credible.2) retrieval model there are certain errors, rather than a true reflection of the people to find information on the problem of the essence. Is not obtained by the retrieval model, the results are credible. The field of information retrieval research has several well-known retrieval model based on the link to the various methods, such as the vector space model, probabilistic model, language model, and Web search, in line with the above two premises.

Based on these two assumptions, we propose the following two fitness function:

$$funtion_1 = \sum_{i=1}^{N} \sqrt{\frac{1}{M} \sum_{j=1}^{M} (s_{i,j} - \overline{s}_i)^2} \quad (2)$$

$$funtion_2 = (\sum_{i=1}^{N} \sum_{j=1}^{M} s_{i,j}) / (N \cdot M) \quad (3)$$

Given in equation (2) the distinction between assumptions based on the credibility of the assumptions and the search results, the variance of the first M N users query a search articles results document similarity score. It represents the search results to distinguish between the degree of similarity score with a clear distinction between the retrieved document similarity variance, the better the current system state; the contrary, the current system is not in the best condition, should be the parameter adjustment.

Formula (3) based on the credibility of the assumptions of the retrieval model is the average of N users query a search before M chapter the results of document similarity. The greater the similarity of the mean, this part of the document and query, the stronger the results more credible, the corresponding system parameters are better.

4 Experimental results

4.1 Comparison with the best performance

To compare system performances, and study the performance of randomly selected parameters of the upper two sets of results. Randomly selected parameters within the meaningful range of model parameters for each user query, randomly generated 100 sets of parameter values, the average of the performance of the system search.. Figure 4 and Figure 5 shows the effect of parameter learning, respectively, MAP and avg (10) as a system performance evaluation.

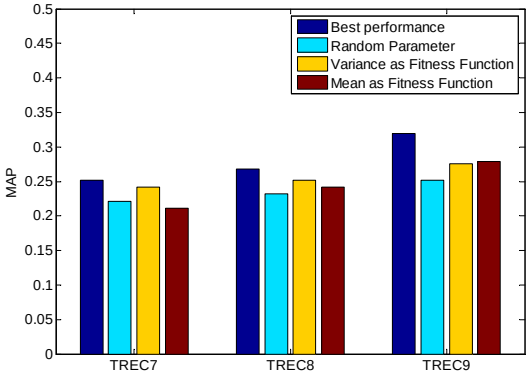


Fig. 4 Effects of parameter learning (MAP)

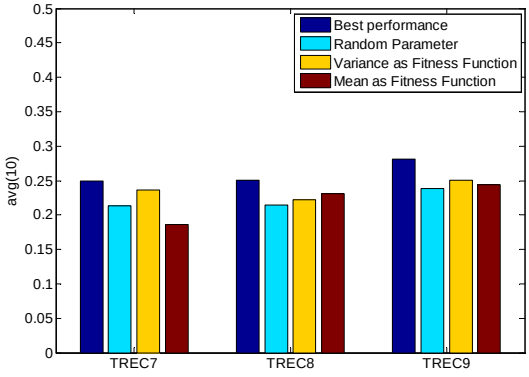


Fig. 5 Effects of parameter learning (avg(10))

Respectively, compared random set of parameter values, the maximum system performance, the use of the former M chapter the results of the similarity of the document, and as the fitness function, and the results of first M articles document the similarity variance and as a fitness function results (N= 100). As shown in Table 4, Table 5 is the first 10

documents retrieved avg (10) the effectiveness of learning for the performance evaluation parameters and retrieval system, Figure 4 corresponds to the MAP.

Table 4 Effects of parameter learning (MAP)

Method	TREC7	TREC8	TREC9
1)Random	0.2211	0.2315	0.2518
2)Best	0.2517	0.2683	0.3189
3)Variance	0.2421	0.2515	0.2756
3) vs. 1)	+9.5%	+8.6%	+9.5%

Table 5 Effects of parameter learning (avg (10))

Method	TREC7	TREC8	TREC9
1)Random	0.2103	0.2145	0.2390
2)Best	0.2489	0.2503	0.2813
3)Variance	0.2362	0.2222	0.2503
3) vs. 1)	+12.3%	+3.6%	+4.7%

You can see, the similarity of the first M documents and as the fitness function can only be achieved good results in the TREC7 and TREC8 on in TREC9 data sets, The avg (10) is much lower than the first M documents similarity variance as to adapt to the effect of the function. While the latter showed a powerful system to retrieve, in most cases are very close to the maximum performance even up to learn.

4.2 Comparison with traditional methods

In fact, the search does not know each query document sets the upper limit of learning is only an ideal situation. Fairer comparison is to the effect of traditional methods to make the system performance in a data set to achieve the optimal parameters as experience in the other two data sets to examine the system performance, and parameter learning for comparison. Table 6 shows the results of the MAP as the evaluation of the performance. Set parameters in accordance with experience, the first data set in TREC7 a set of optimal parameters, and this set of parameters for TREC8, and TREC9 of data to get the system performance. The other hand, the learning method uses dynamic parameters, respectively, on three different test data sets, using a genetic algorithm for the automatic guidance of dynamic parameter learning, which before the definition of similarity variance of 100 returned results for the fitness function.

Table 6 Effects of parameter learning (MAP)

Method	TREC7	TREC8	TREC9
1)Using best parameter in TREC 7	0.1679	0.1876	0.2016
2)GA based dynamic parameter learning	0.1987	0.2087	0.2311
2) vs. 1)	+18.3%	+11.2%	+14.6%

Can be seen, compared with the traditional experience the value of setting parameters, dynamic parameters of the proposed learning method on the system performance in TREC7 dataset is very close to optimal results, and at the same time, respectively, to bring in TREC8, and TREC9 Data 11.2% and 14.6%, a substantial increase system performance.

5 Conclusions

In previous retrieval models, the parameter set is a very important issue, which largely determines the retrieval performance. Traditional methods based on the experience to set system parameters by hand. In this paper, we first take an experiment to investigate the generalization ability of the experience set the parameters on the large-scale collection of standard test to get:

1) As different data collection be retrieved using the same system, the optimal performance parameters may be completely different; 2) even if the data set is the same user query, retrieve the best performance parameter values may be different; 3)Even if you use the same set of data and user queries, but to evaluate the retrieval performance of different standards, different optimal parameters.

In addition, due to lack of information retrieval, it can not be guided learning. In this paper, the retrieval model is firstly proposed, on basis of which, the credibility of the assumptions and retrieve the results of the distinction between assumptions are built. Then, we proposed a similar relationship between the two methods, which are based on document and query the fitness function: to retrieve the results of the first M documents the similarity of the mean for the first time and similarity variance. Then by using the genetic algorithm, given a retrieval system based on unsupervised learning parameters dynamically adjust system

parameters automatically learning the problem is resolved.

Experiments show that the similarity varies as the fitness function. With our method, system performance is always close to or even reach the upper limit of the model may achieve optimal performance based on genetic algorithm parameters dynamically learning; with the traditional experience of the value of parameter set given method compared to a very substantial increase system performance, which is a very effective solution.

Future studies will examine more of the fitness function and to study other than the genetic algorithm learning speed faster, more efficient machine learning methods for retrieval system parameter learning.

Reference

[1] Jinyoung Kim, Xiaobing Xue and W. Bruce Croft，A Probabilistic Retrieval Model for Semistructured Data，Lecture Notes in Computer Science, 2009, Volume 5478/2009, 228-239

[2] Hammouda, K.M., Kamel, M.S., Efficient phrase-based document indexing for Web document clustering, Knowledge and Data Engineering, Volume: 16 Issue: 10， 2004，pp：1279 – 1296

[3] Stephen Robertson, Hugo Zaragoza, Michael Taylor, Simple BM25 extension to multiple weighted fields, Proceeding CIKM '04 Proceedings of the thirteenth ACM international conference on Information and knowledge management, 2004

[4] E. M. Voorhees, D. Harman, Overview of TREC2001, Proc. of the 10th text retrieval conf. gaithersburg, MD:NIST Special Publication, 2011, 1-15

[5] Weiss, T., Hillenbrand, J., Krohn, A.,Jondral, F.K., Mutual interference in OFDM-based spectrum pooling systems, Vehicular Technology Conference, 2004, vol. 4, pp:1873 – 1877

[6] Deb, K., Pratap, A., Agarwal, S., Meyarivan, T., A fast and elitist multiobjective genetic algorithm: NSGA-II, IEEE Transactions on Evolutionary Computation, Volume: 6 Issue: 2, 2002, pp: 182 – 197

[7] J. Holland, Adaption in Natural and Artificial Systems. Ann Arbor, Michigan: University of Michigan Press, 1975

[8] Kazarlis, S.A., Papadakis, S.E., Theocharis, J.B., Petridis, V.,Microgenetic algorithms as generalized hill-climbing operators for GA optimization, IEEE Transactions on Evolutionary Computation,Volume: 5 Issue:3 ,2002, pp:204 – 217

[9] Dana Vrajitoru, Crossover improvement for the geneticalgorithm in informationretrieval, Information Processing & Management,Volume 34, Issue 4, July 1998, Pages 405–415

[10] D. Vrajitoru, Crossover improvement for the genetic algorithm in information retrieval, Information Processing and Management, 1998, 34(4): 405-415

XML Hash File: A new XML Storage Structure for the Efficient Processing of Path Queries

Muhammed Al-Mulhem [*1,*], Salahadin Mohammed [†1], and Mahboub A. M. Naji [‡1]

[1]*Department of Information and Computer Science, KFUPM*

ABSTRACT

Many XML storage structures have been proposed in the literature. Most of these storage structures result in high query cost because of the way they organize the XML data. In this paper, we propose a new XML storage structure known as *XML Hash File* (XHF). The XHF encodes, organizes, and partitions XML data in a way that results in a lower query cost and without compromising disk space. Based on XHF, we also propose a new path query processing algorithm known as XHF-Path. We ran many experiments to compare XHF-Path with TwigVersion, one of the best XML query processing algorithm. Our experimental results confirm that XHF-Path has significantly better performance than TwigVersion.

KEY WORDS: XML, XML indexing, XML storage, XML query processing, XML database, Twig pattern matching, Storage structure.

1 Introduction

The exponential growth in the use of XML has heightened the need to efficiently store large volumes of XML data. This is because, the way XML data is organized in a secondary storage has significant effect on the cost of storing and querying the data. Many structures of storing XML data have been proposed in the literature [4, 6]. Most of these structures organize XML data in a way that results in a higher query cost.

An XML document can be represented as an ordered, labeled, and rooted tree know as *XML-tree* [11]. Each XML-tree node corresponds to an XML element or an attribute or a value. An edge of an XML-tree represents a parent-child relationship between two nodes. Querying XML data involves searching for nodes and structural relationships that satisfy the constraints specified by the query. Common XML query languages, such as XPath and XQuery, specify patterns of selection predicates on multiple nodes which have the required structural relationships. One of the most common XML queries is path query. Finding all the matches of a path query is the core operation of XML query processing. Many path queries have been proposed in the literature [1, 2, 4, 5, 8–10, 13, 14]. Most of the existing path queries result in a high query cost because: (i) they are based on inefficient storage structures; (ii) they don't have the ability to skip elements that don't contribute to the final answer of a query; (iii) they generate intermediate results more than is necessary; and (iv) they perform many unnecessary join operations. In this paper we propose a path query processing algorithms know as XHF-Path. The XHF-Path algorithm is based on the proposed storage structure, XHF. They address all the above mentioned limitations of the existing algorithms. We conducted several experiments to compare the performance of XHF-Path with that of TwigVersion, one of the best query processing algorithms [14]. The experimental results show that XHF-Path has significantly better performance than TwigVersion.

The rest of this paper is organized as follows. XHF and XHF-Path are discussed in sections 2 and 3 respectively. Experimental results of XHF-Path are presented in Section 4. Section 5 is the conclusion.

2 The XML Hash File (XHF)

This section explains the proposed XML storage structure in detail. But before we explain the proposed XML storage structure, let us first define some terms used in the rest of this paper.

[*]mulhem@kfupm.edu.sa: To whom correspondence should be addressed.

[†]adam@kfupm.edu.sa

[‡]te_mahboob@yahoo.com

2.1 Definitions

Definition 2.1 ***node-name****: is an element name or an attribute name. For example, library, book, and title are node-names.*

Definition 2.2 ***node-label****: is a label assigned to a node by a labeling scheme. For example, Dewey Ids.*

Definition 2.3 ***root-path****: is a path from the root node to any node.*

Definition 2.4 ***label-path****: is the ordered list of slash separated node-names of a root-path.*

2.2 The XHF Main Structures

The XHF consists of the following four main structures.

1. ***NBS-table:*** The XHF, using a hash function, Φ, assigns each distinct node-name in a level of an XML-tree a unique bit string which we call *Node Bit String* (NBS).

$$NBS = \Phi(n_i) \quad (1)$$

where Φ is a hash function and n_i is a node-name at level i of the XML-tree. NBS-table is a table which contains the list of node-names and their corresponding NBSs. For example, the table in Figure 2a shows the NBS-table of the XML-tree shown in Figure 1.

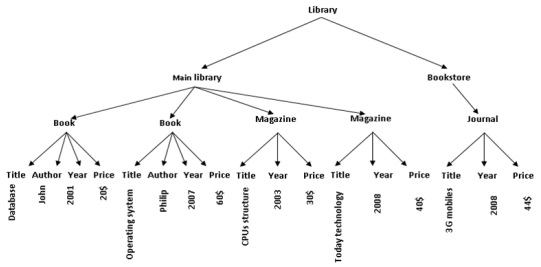


Figure 1: Sample XML tree

Node-level	Node-name	NBS
1	Main library	0
1	Bookstore	1
2	book	01
2	Magazine	10
2	Journal	11
3	Title	001

(a) Sample NBS table

Label-path	PBS
/Library/Main library	0
/Library/Bookstore	1
/Library/Main Library/book	001
/Library/Main Library/Magazine	010
/Library/Bookstore/Journal	111
/Library/Main Library/book/title	001001

(b) Sample PBS table

Figure 2: Sample NBS and PBS tables

2. ***DV-table:*** Each record of the DV-table, DV-record, contains the Dewey-ID of the last node of a root-path. If the last node of a root-path is a leaf node, then the DV-record contains both the value and the Dewey-ID of the node. DV-records of identical root-paths are stored in the same disk blocks. Let us call a disk block which contains DV-records as a DV-block. DV-table is implemented as a tree, DV-tree. We implemented the DV-tree of a non-leaf node as a B+-tee and that of a leaf-node as a BANG file [3]. DV-tree of a leaf-node can also be implemented using any multidimensional file structure. XHF has one DV-tree for each distinct label-path; thus, the number of DV-trees in a XHF is equal to the number of distinct root-path names in the XML dataset.

3. ***PBS-table:*** The XHF labels each label-path in an XML-tree a bit string called *Path Bit String* (PBS) using a hash function.

$$PBS = \Psi(p) = \Phi(p_1)||\ldots||\Phi(p_k) \quad (2)$$

where Ψ is a hash function, p is a label-path, p_i is a node-name at level i of p, $||$ is a concatination operation, and k is the length of p. Inshort a PBS of a root-path is the ordered concatenation of its NBSs.

The table in Figure 2b shows label-paths and their PBSs. Each PBS-table record (PBS-record) contains a PBS and the address of its corresponding DV-tree. The number of PBS-records in a PBS-table is equal to the number of distinct PBSs (distinct label-paths) in the XML-tree. Let us call a disk block which contains PBS-records as a PBS-block. When the number of PBS-records in a PBS-block exceeds its capacity, the PBS-records are split between two PBS-blocks. The split is based on the value of a PBS bit position called *splitting-bit*. Let PBS_i be the i^{th} bit (from the left) of a PBS. The splitting-bit is the left most bit were the PBSs of a PBS-block differ. For example, if all the PBSs in PBS-block have the same PBS_1 value, the same PBS_2 value, but different PBS_3 value, then the splitting-bit is PBS_3. Those PBS-records with splitting-bit value of 0 will be stored in one PBS-block and those with splitting-bit value of 1 will be stored in another PBS-block. If a splitting-bit doesn't result on an even split of the PBS-records between the two PBS-blocks, then the next splitting-bit is used to split the records; and if this one also doesn't result on an even split, then the next one is considered and so on. This splitting strategy will reduce the number of splits and will

increase the average number of records per block specially when the PBSs are skewed.

4. ***XHF-directory:*** Each XHF-directory record (dir-record) consists of a key (a bit string), the number of bits in the key, and a PBS-block address. The number of bits in the key is equal to the position of the splitting-bit that was used to create the corresponding PBS-block. For example, if PBS_j was used as a splitting-bit when creating a certain PBS-block, then the number of bits in its corresponding key is j and the key is $PBS_1 PBS_2 \ldots PBS_j$.

   When a PBS-block which was created using PBS_j is full, it is split using a splitting-bit PBS_{j+k}, where $k > 0$. If $k = 1$, then the length of the key of each of the two new PBS-blocks is $j + 1$; but if $k > 1$, then the length of the key of the PBS-block whose splitting-bit value is 0 will remain as j and the length of key of the second PBS-block will be $j + k$.

   Let us call a disk block which contains dir-records as a dir-block. When a dir-block is full, it is split in the same way a PBS-block is split. Splitting the root dir-block results in three new dir-blocks. Two of these dir-blocks will split the contents of the old root, and the third dir-block will be the new root. The new root will contain two dir-records, one for each of the new dir-blocks.

2.3 Searching the XHF

In XHF, searching for a path or a node specified by a query starts by mapping the specified path or node into all possible label-paths. Then the search for each label-path starts from the root dir-block of the XHF. All the keys in the root dir-block are matched with the PBS of the label-path. They shortest key that entirely matches the whole PBS of the root-path name is selected. Then the search moves to the dir-block of the selected key. The same process is repeated in each selected dir-block until the search reaches a PBS-block. If the PBS that we are searching is in the PBS-block, then all the Dewey-IDs in the corresponding DV-tree are selected. If the search label-path includes a specified value, then only Dewey-IDs associated with values matching the specified value are selected.

3 The XHF-Path Algorithm

The pseudocode of XHF-Path is shown in Algorithm 1. To answer a simple path query, XHF-Path first maps the simple path query into all possible label-paths (Line 1). Then it converts each label-path into its equivalent PBS using the hash function Ψ of Equation 2 (Line 3). Then its searches for the DV-tree of each PBS using a search algorithm similar to the one described in Subsection 2.3. The search algorithm returns the address of the corresponding DV-tree (Line 4). If the query specifies a value predicate, then the Dewey-IDs associated with values matching the specified value are returned from the DV-tree as output (Line 5). If the query has no value predicate, then XHF-Path returns all the Dewey-IDs stored in the DV-tree (Line 5).

Algorithm 1: XHF-Path

```
1 set(path-name) ⇐ Derive-path-names(Q);
2 foreach p IN set(path-name) do
3     PBS ⇐ Ψ(p);
4     DV-tree ⇐ Find-DV-Tree(PBS);
5     Retrieve(DV-tree);
6 end
```

4 Performance Analysis

4.1 Experimental setup

The machine: The proposed algorithms were implemented on a laptop computer with a 1.7 GHz Dual Core processor and a 1 GB of RAM and running Windows XP.

In order to study the performance of XHF-Path, three benchmark datasets, namely, DBLP, SwissPort, and XMark, were used in the experiments [12]. We also used several queries with different properties, such as, axis types, selectivity ratio, output size, etc. as shown in Figure 3. To measure the performance of the proposed algorithms, we used the three commonly used parameters, namely, *I/O cost* (number of elements read), *Elapsed time*, and *Storage cost*.

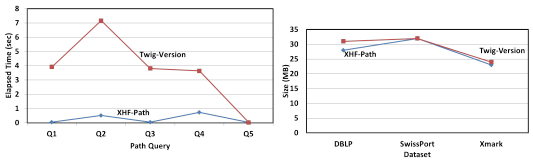
#	Path Query
Q_1	/dblp/article[title="Principles of Distributed Object Database Languages."]
Q_2	/dblp/inproceedings[year="2002"]
Q_3	//inproceedings[crossref="conf/performance/2000"]
Q_4	//article[.//number="1"]
Q_5	//mastersthesis[school="University of Texas, Austin"]

Figure 3: Path queries used in the experiments

4.2 The performance of XHF-Path

In this section, we compare the performance of XHF-Path with that of TwigVersion using the queries shown in Figure 3 and the DBLP dataset. The elapsed time for

processing these queries using XHF-Path and TwigVersion are shown in Figure 4a. This time is mainly due to I/O cost because both XHF and TwigVersion store XML data path-based, hence no join operation is needed. We can see that XHF-Path consistently outperforms TwigVersion, except for Q_5. To understand the reason behind the superiority of XHF-Path over TwigVersion consider query Q_1. The value node *"Principles of Distributed Object Database Languages"* has high selectivity, it appears only once at the end of the document. XHF-Path processes highly selective queries efficiently since only few blocks in the corresponding V-Tree are accessed. On the other hand, TwigVersion has to examine all the elements in the *title* data cluster to retrieve all *title* elements with the value *"Principles of Distributed Object Database Languages"*. For Query Q_5, even though the value node *"University of Texas, Austin"* has high selectivity, the path *"/mastersthesis/school"* is also highly selective, only five elements match this path. These elements are clustered in one disk block in both TwigVersion and XHF. As a result the performance of XHF-Path and TwigVersion were similar. Figure 4a shows that XHF-Path is up to 91 times faster than TwigVersion for highly selective queries.



(a) Elapsed time (b) Storage size

Figure 4: Elapsed time and storage cost

4.3 XHF storage space

The widely used file structure to store XML datasets in native XML databases is node-based inverted-list. An inverted-list is created for each distinct node-name. The XHF storage structure is label-path based. In the XHF values and Dewey-IDs belonging to identical root-path names are stored in the same BANG file. Figure 4b shows the amount of disk space used by the XHF and the XISS to store the same XML dataset. XISS uses element-based inverted-lists [7]. As can be seen form the figure, both storage structures used nearly the same amount of disk space to store the same XML dataset. This confirms that no storage space was compromised by the XHF in order to improve query performance.

5 Conclusion

The way XML data is organized in a secondary storage affects the cost of querying the data significantly. In this paper we propose a new XML storage structure, known as XHF, which organizes XML data in a way that minimizes query cost and without compromising disk space. Unlike node-based inverted-list, XHF stores a path as it is without shredding it into its component nodes and then storing each node in different inverted-list. Based on XHF, we also propose a new path query algorithm called XHF-Path. We run many experiments to compare XHF-Path with TwigVersion, one of the best XML query processing algorithms, using both synthetic and real XML datasets and different types of queries. XHF-Path showed significantly better performance in almost all the experiments.

Scope of future work includes extending XHF-Path to handle twig queries with all kinds of structural relationships (such as, following sibling, preceding sibling, and so on) and all types of value constraints (such as, inequalities and all types of logical operators).

Acknowledgements: The authors are grateful for the support provided by King Fahd University of Petroleum and Mineral, Dhahran, Saudi Arabia.

References

[1] Shurug Al-Khalifa, H. V. Jagadish, Nick Koudas, Jignesh M. Patel, Divesh Srivastava, and Yuqing Wu. Structural joins: A primitive for efficient xml query pattern matching. In *Proceedings of the 18th International Conference on Data Engineering*, ICDE '02, pages 141–, Washington, DC, USA, 2002. IEEE Computer Society.

[2] Nicolas Bruno, Nick Koudas, and Divesh Srivastava. Holistic twig joins: optimal xml pattern matching. In *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, SIGMOD '02, pages 310–321, New York, NY, USA, 2002. ACM.

[3] M. Freeston. The bang file: A new kind of grid file. In *ACM SIGMOD Int. Conf. on Management of Data*, 1987.

[4] Gang Gou and Rada Chirkova. Efficiently querying large xml data repositories: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19:1381–1403, 2007.

[5] Nils Grimsmo, Truls A. Bjørklund, and Magnus Lie Hetland. Fast optimal twig joins. *Proc. VLDB Endow.*, 3:894–905, September 2010.

[6] S. Haw, , and C. Lee. Data storage practices and query processing in xml databases: A survey. *Knowledge-Based Systems*, 24(8):1317–1340, 2011.

[7] Quanzhong Li and Bongki Moon. Indexing and querying xml data for regular path expressions. In *IN VLDB*, pages 361–370, 2001.

[8] Jiaheng Lu, Tok Wang Ling, Chee-Yong Chan, and Ting Chen. From region encoding to extended dewey: on efficient processing of xml twig pattern matching. In *Proceedings of the 31st international conference on Very large data bases*, VLDB '05, pages 193–204. VLDB Endowment, 2005.

[9] Jiaheng Lu, Xiaofeng Meng, and Tok Wang Ling. Indexing and querying xml using extended dewey labeling scheme. *Data Knowl. Eng.*, 70:35–59, January 2011.

[10] Praveen Rao and Bongki Moon. Prix: Indexing and querying xml using prüfer sequences. In *Proceedings of the 20th International Conference on Data Engineering*, ICDE '04, pages 288–, Washington, DC, USA, 2004. IEEE Computer Society.

[11] Igor Tatarinov, Stratis D. Viglas, Kevin Beyer, Jayavel Shanmugasundaram, Eugene Shekita, and Chun Zhang. Storing and querying ordered xml using a relational database system. In *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, SIGMOD '02, pages 204–215, New York, NY, USA, 2002. ACM.

[12] http://www.cs.washington.edu/research/xmldatasets.

[13] Haixun Wang, Sanghyun Park, Wei Fan, and Philip S. Yu. Vist: a dynamic index method for querying xml data by tree structures. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, SIGMOD '03, pages 110–121, New York, NY, USA, 2003. ACM.

[14] Xin Wu and Guiquan Liu. Xml twig pattern matching using version tree. *Data Knowledge Engineering*, 64(3):580 – 599, 2008.

SESSION

KNOWLEDGE AND INFORMATION MANAGEMENT AND ENGINEERING

Chair(s)

TBA

Digital Interactive Public Pinboards for Disaster and Crisis Management - Concept and Prototype Design

[1]Peter-Scott Olech, [1]Daniel Cernea, [1]Helge Meyer, [1]Sebastian Schoeffel, and [1]Achim Ebert
[1]Computer Graphics and HCI Group, University of Kaiserslautern, Kaiserslautern, Germany

Abstract— *Recent natural disasters, like the earthquakes in Port-au-Prince, Haiti (2010), Christchurch, New Zealand (2011) and the Tohoku earthquake in Japan (2011), which also triggered a tsunami resulting in the catastrophic failure of numerous nuclear power plants, pose the question how to support first responders in providing fast and adequate help. When first responders arrive on-site it is crucial that the flow of information is ensured: important fields are logistics, communication, personal management and the deployment of up-to-date information. Unfortunately the events in the past showed that there are certain shortcomings, especially in terms of communicating information from local responders to arriving responders. Our approach proposes the utilization of large public displays, seizing the idea of traditional pinboards, referred in our work as Digital Interactive Public Pinboards (DIPP). DIPPs are set up in hot spot locations and provide fast and reliable information to first responders as well as citizens in the area of the natural disaster.*

Keywords: Large displays, Public Displays, Mobile Interaction, Information Hotspot, and Decision Support.

1. Introduction

The years 2010 and 2011 drastically showed how vulnerable we are when it comes to natural disasters. Therefore it is essential to provide effective help as fast as possible. After recent natural disasters, like the earthquakes of Port-au-Prince, Haiti (2010) and earthquake of Tohoku, triggering a tsunami and causing a nuclear disaster, we learned that international aid can be improved. For example, responders in Haiti had to struggle with outdated road maps; all while roads where impossible to pass due to debris. Precious time has been lost in order to find an alternate route to arrive the destination location on-site, where help was needed so urgently. Another example of disaster management gone bad was the way it was dealt with foreign first responders after the earthquake/tsunami disaster in Japan in 2011: arriving first responders had to come off empty handed, shortly after arrival, due to a serious lack of information and coordination.

Providing up-to-date information, ensuring constant information flow and reaching the people, in need of this information, to organize and coordinate, are the designated factors to ensure efficient help, aside from technical factors. In this paper we address existing issues by introducing *Digital Interactive Public Pinboard*, as an efficient and reliable way to share information in case of natural disasters. Traditional bulletin boards have been used to post messages like advertisements, public announcements and lost & found messages in public. The chosen locations of such traditional bulletin boards were places which were frequently visited, by a large number of people, namely *hotspots*.

However, the problems of these traditional bulletin boards were the correctness as well as actuality of the posted information. In order to optimize this, we propose DIPPs, set up in hotspot locations. The advantage of hotspot locations is the availability to a large number of people, like first responders arriving on airport and immediately having access to an up-to-date situation report with important information like current news, where to find shelter, and where is help needed. Furthermore another issue is addressed by using DIPP in hotspot locations: having a fixed location for the public displays, it is obvious for users where they can access information. DIPPs also serve as contact points, enabling social interaction between first responders.

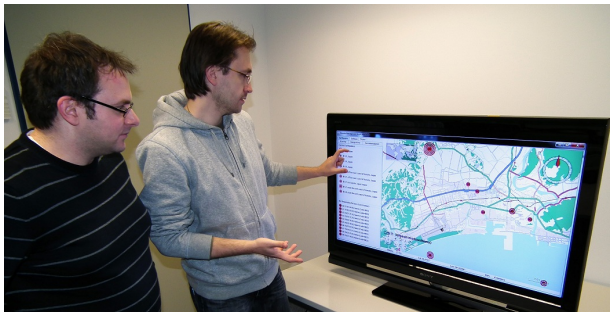


Fig. 1: Testing the prototype on a large display.

An important requirement for DIPPs is comprehensibility of the provided information. A variety of users, with diverse backgrounds, have to be able to access information and to understand the displayed content. Therefore the visual representation of the displayed information has to be easy accessible and comprehensible without a long training period and studying a user guide. Since smart phones and handheld devices are very popular today, new ways of interaction are enabled, due to the functionality these devices offer, nowadays. Modern mobile phones allow for many communication modalities: WIFI network, Bluetooth, GPS, camera etc. which can be exploited in natural disaster scenarios.

The experiences of the Tohoku earthquake and the aftermath showed that Internet still was available, when landline, as well as mobile phone networks, has been compromised.

The main goal of our approach is to overcome the shortcomings of past experiences in ensuring an adequate flow of information tailored to our DIPP approach, featuring a system able to provide the needed information to the users by mobile device interaction, to achieve maximum flexibility.

In the following we will provide a brief overview of related work, especially in the fields of public displays and also disaster management. Then we provide a more detailed view of DIPP, describing the concept. After that a system prototype is described, giving an overview of the client/server architecture. Finally we highlight the conclusion and outlook for this research.

2. Related work

Using public displays to present news and advertisements has been popular for years. Digital billboards mainly have been used for signage and for commercial advertisements; users could watch but interaction was limited or impossible. Users today have a different mindset: they expect being able to interact with displays. In the following we provide a brief overview about recent developments in the area of public displays and digital billboards.

The *Notification Collage* (NC) is introduced by Greenberg and Rounding [Greenberg]. With this groupware system users are enabled to share files and information on a designated real-time collaborative display. Within an office environment sharing of files is made possible by the NC, which basically is similar to a public bulletin board. Lui et al. [15] present an *Interactive Wireless Electronic Billboard*. With their approach users are able to see advertisements on public displays and interact with the display. With mobile phone interaction the user is able to retrieve further product information and even purchase the advertised product. *Plasma Posters*, an approach of Churchill et al. [5], are used to share information within teams and different groups of users. These *Plasma Posters* are used in an office environment to share information in public spaces. The approach itself is described and an evaluation is conducted. The *KIMONO* (Kiosk mobile-phone knowledge sharing center) allows user to interact with a public display via mobile phones. Both acquisition and exchange of data are enabled [11]. Gronbak at al. introduce *InfoGallery* [9], an approach to provide visitor guides, visitor information and other digital content via public screens in libraries and art galleries. With *BlogWall* [3] Cheok et al. present an approach to use mobile messages displayed on public screens for artistic and social communication. With the *BlogWall* approach a new form of cultural computing is created, because it also features multi-language support. Hosio et al. [10] present the utilization of public displays and mobile phone interaction in an urban environment. The approach is also evaluated and limitations are discussed. Thelen et al. present a *Digital Interactive Pinboard* [23], enabling users to interact with public displays. The public screen is used as a digital pinboard, enabling users to share multimedia information by mobile phone interaction. Two scenarios are presented and discussed: an office and an enterprise scenario.

The discussion of how to harness information technology to respond and to manage natural disasters, as well as man-made disasters, is going on for years. Quarantelli [20] examines ten non-technical questions, dealing with the issues of information overload, communication between different actors involved in disaster response, as well as the crucial problem of information validity, if the information provided is current or already outdated. Stephenson and Anderson [22] give a historical overview in how information technology (IT) and computer science have been used from the 1970s and also provide an outlook of future utilization possibilities in the field of disaster management. Fischer [7] also provides an overview of new information technologies and how they have been used to help in emergency mitigation and also how these technologies can be used in enhancing training scenarios for emergency personnel. The work of Kapucu [13] examines the role of interagency communication networks during emergencies. He presents a theoretical framework and illuminates the positive as well as the negative aspects of information technologies and interagency communication. Marincioni [16] proposes that information technology has to reach one level, so that the diverse actors involved in disaster management can equally use the available knowledge and information. Mes [17] examines in how population in an earthquake zone can be warned in a most efficient way. He points out that mobile phone networks tend to survive earthquakes, which is true for the Haiti earthquake (2010) but not valid for the Japan earthquake (2011) where both landline and mobile phone networks broke down, making communication over phone impossible. The article of Underwood [24] points out that the sheer availability of technologies (both hardware and software) do not resolve crisis. They have to be adopted by governments as well as aid agencies. Landgren and Nulden [14] present how to use patterns of mobile phone interaction for organizing time-critical data in the field of emergency response. Disaster education with the *Sandankai System Method* [18] is performed in a workshop environment, in order to develop an action plan in case of a disaster. Na et al. carry out a case study in a high school, in order to develop a three-step action plan to respond to a possible disaster. In the work of Palen et al. [19] an outlook of future emergency management is given. Rauschert et al. [21] propose the utilization of large screen displays and novel interaction methods (voice and gesture recognition) to overcome the shortcomings of traditional Geographic Information Systems (GIS), allowing multi-user interaction during emergency management situations. Currion at al. introduce open source software for disaster management

[6]. Their *Sahana* disaster information system is evaluated. Another open source approach is presented in the work of Abed et al. [1]. The aim of the work is the study on how open source web based GIS can be used to aid in emergency response and to which level it is a suitable tool for decision making. Chen et al. [4] propose the implementation of GIS and computational models into the emergency management and spatial decision making process.

3. Conceptional Design

In the following subsections we will provide an overview on what we considered in the development of the concept and also implementation of a first prototype of our disaster and crisis management system.

3.1 General considerations

According the definition of disaster, it is a damaging event, that can't be resolved neither at a local nor at a supra-regional level within a reasonable timeframe. Therefore exterior levels have to be included in order to resolve this damaging event, which in general means international aid [2]. Having that in mind we did focus on external first responders, arriving on-site and needing information and guidance. The Haiti earthquake (2010) has shown that not only large organizations like the Red Cross are providing aid (humanitarian aid, search and rescue etc.), but also smaller organizations (Non-government organizations, NGOs) or even individuals are reaching disaster areas in order to provide help. Our goal was to provide a centralized information platform that captures current information for all help parties, in order to organize arriving responders and provide them with all necessary information.

3.2 Why Public Displays?

Inspired by traditional pinboards, which are still well known from universities and supermarkets, we want to provide arriving aid personnel with a contact point at a so called hotspot location. The modern day variant of a pinboard features a large public display, where basically all important information is displayed, visualized in a way that allows groups with a diverse background to perceive and comprehend the data fast fast, without a long training period. Effectiveness and efficiency is key, to minimize losses and provide suitable help as fast as possible.

3.3 What is the purpose?

The purpose of our system is to provide a basic, locally centralized information system to organize, collect, share, and distribute information to first responders after natural disasters. Therefore we propose the use of digital interactive public pinboards enabling user interaction with mobile devices over WIFI network. Our system is a client/server system, providing an information infrastructure over a WIFI network at hotspot locations. The clients (mobile phones) can communicate with the server (for example registration, uploading pictures, and search functionality).

3.4 What are the essential features needed?

Considering the previous subsection the following features are essential for arriving first responders:

First responder management: by creating a user profile (name, profession, contact information like email address, GPS camp coordinates etc.) users are able to search for other first responder e.g. by profession in order to collaborate and in this way facilitate the cooperation and being able to ask for support.

Deployment of up-to-date information: up-to-date information can be accessed at the information hotspots (e.g. map material, actual satellite images, contact information, GPS coordinates of accommodation possibilities, camps, where help is needed). The client/server approach also minimizes the risk of unverified and invalid information.

Logistics: Organization and coordination of accommodation, water supply, and food supply, medical supplies, coordination of search & rescue teams etc.

Communication: Centralized communication, to ensure that the information provided to first responders is accurate, up-to-date, and verified.

Hazards: Visualization of known hazards, e.g. aftershocks, leaking oil pipelines etc.

3.5 What features should be included?

Besides the essential features, there are many other features that are useful and can extend the user group also to the population in the disaster area.

Missing people: The functionality of adding missing people and providing basic information (name, picture, last known position).

Specific tasks: organize open tasks: search for missing persons, explore area which has been damaged badly, restore basic infrastructure. If an engineer is taking care of a specific task, the engineer marks the task as *taken*, in order to ensure that there is no overlap on tasks (Open task, Pending/taken task, and Solved task). Furthermore information about the group or individual, the previous record in solving tasks is displayed (competence level, reputation).

Visual representation of information: the representations will not only include statistics, but also predictions (e.g. water supply sufficient for 2 more days). The users can

interact, collaboration in the UI and in the visualization data has to be supported, most likely asynchronously.

4. The Prototype

The prototype of the DIPP system is based on the NASA World Wind Java SDK [1]. Based on the conception of an emergency management system, described in the previous section, a prototype is implemented, realizing these requirements. The NASA WWJ SDK framework was released in July 2011 as version 1.2, for the first time as a *stable release*. The framework is flexible, so that it can be extended and tailored to the needs of the user. Similar to Google Earth [2] data is loaded from local buffer memory and a 3D view is offered, based on a peculiar elevation model. Unlike Google Earth it offers the advantage of being platform independent; if the OS supports OpenGL, one is good to go. Another plus is that the framework is published under NOSA license (internal NASA licensing, Nasa Open Source Agreement), making it available for free to the research community.

4.1 Implementing the server

One of the key features of our prototype is accessibility and therefore the visual representation of events. For representing the events on the DIPP we did choose symbols which are easily recognized, without users having to read a manual or go through a list of keys. To ensure a high recognition value we designed the symbols according to the standards of U.S. Department of Defense (DoD).

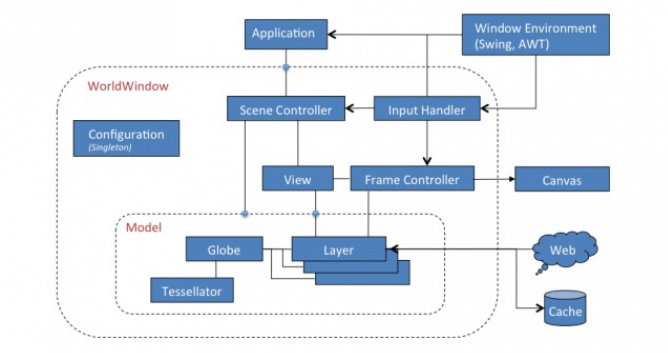


Fig. 2: WWJ framework overview (goworldwind.org).

The basis of our application is map material from the Open-Street Map (OSM) project [3]. It is editable and can be provided from designated servers. In addition it is the only map material which is available for free and offers high precision. GeoRSS offers additional geo data tailored to specific application areas, for example earthquake data is provided by the United States Geological Survey (USGS).

[1] http://worldwind.arc.nasa.gov/java/
[2] http://www.google.com/intl/en/earth/index.html
[3] http://www.osmfoundation.org

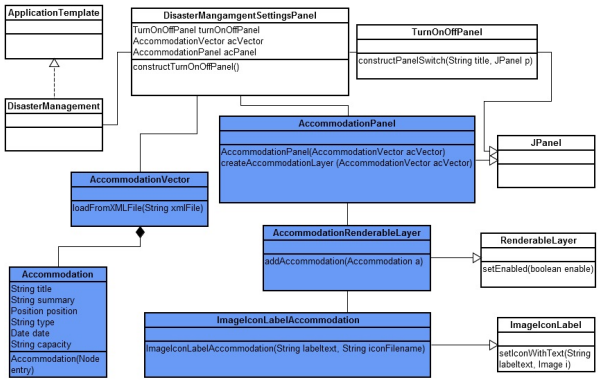


Fig. 3: Prototype implementation. Accommodation class.

Figure 2 provides an overview of the framework. The globe model represents a planet and an elevation model. A *tesselator* generates the elevation model and layers are projected onto the globe, displaying grid and vector data. The displayed objects keep their position during user interaction (navigation). The model consists of all data responsible for displaying the layers. The view sets the user's view on the model, triggered by the user and the *InputHandler*. The *SceneController* draws the model and determines when it is drawn, combining view and model. The object *WorldWindow* is produced during run time in the AWT/Swing environment and can be integrated in the Canvas.

In order to implement the requirements for our system prototype, new classes and packages had to be created. Figure 3 shows the 5 standard classes (highlighted in blue) and provides insight of the prototype feature *accommodation*. After creating the *DisasterManagement* main application, a DisasterManagement *SettingsPanel* is created. In this panel the user can make adjustments and enter data, e.g. switch on and off layers. The *DisasterManegementSettingsPanel* creates an accommodation vector by calling the method *loadFromXMLFile* (String xmlFile). After loading the accommodation vector a *AccommodationPanel* is created, which loads the information bar, based on the elements of the accommodation vector. At the same time the objects from the accommodation vector and the method *addAccomodation* (Accomodation a) are parsed to the *Accomodation RenderableLayer*. By doing so an *ImageIconLabelAccomodation* is created and added to the *AccomodationRenderableLayer*. All elements added to the *AccomodationRenderableLayer* are then displayed in the map. The images are loaded from the subfolder *img/*.png*, whereas the file name corresponds the attribute *title* of the XML-file. At the end the *DisasterManagementSettingsPanel* loads a *TurnOnOffPanel* enabling users to turn on or off the different layers. By setting a checkbox the *RenderableLayer* is activated or deactivated with the method setEnabled(boolean enable).

4.2 Features implemented in the server

Based on the requirements posed in section 3, the following features have been implemented in the system prototype.

Fig. 4: Icons. Earthquake. Number of rings represent magnitude according to *Richter* scale (left). Shelter possibilities. Color coding indicating availability (middle). Water supply. Color coding indicating availability (right).

The prototype features an earthquake scenario, as one possible example of natural disasters. Earthquakes are visualized on the map and therefore are visible for both emergency personnel as well as refugees and citizen. The earthquakes are represented as concentric circles, while the number of circle rings provides information on the magnitude. The last aftershock is visualized by a flashing symbol, catching the attention of the user. In addition, the positions of the emergency personnel in the field are displayed, as well as dangerous structures. More details are accessible in the left-hand information bar, providing more in-depth information, if needed.

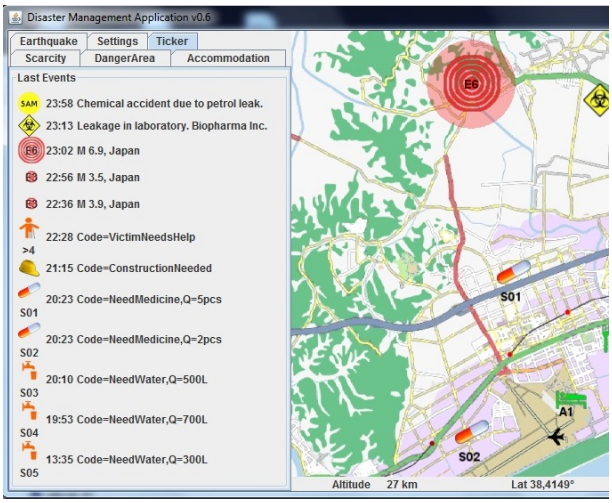


Fig. 5: The News Ticker function allows users to see all events, ordered by time of occurrence, most recent events on top.

Where to go? This crucial question many arriving responders have. Since local personnel is working full capacity to provide help and support for the population in the disaster area, our system can provide guidance to incoming personnel in order to find accommodation. Our system provides information where accommodation is available (green symbol: available; small number, number of available beds; red symbol: occupied) based on the field of expertise. If medical help is needed nearby a certain camp, then a doctor will be advised to go to this camp, by the system (based upon the registration).

The vigilance system informs users about bottlenecks of essential goods, like drinking water, food and medicine. Icons visualize the status of the goods in certain areas. Flashing symbols, showing that e.g. medicine is needed very urgent and where it is needed catch the user's eye. To provide information about hazardous areas the prototype features the functionality to mark potentially hazardous areas, similarly to weather maps. This feature provides refugees information if it is safe to return home or if there are other dangers, like contamination. Emergency personnel in this way can prepare themselves by wearing adequate protective clothing before entering such areas. Our prototype also features news ticker functionality. This function combines the information of all layers and displays everything, unfiltered, based on the time of events. The most recent events are shown on top of the information bar, providing real-time updates of events.

4.3 Features of the client

The client system is implemented in Java and is currently functional under Windows Mobile. Other major mobile platforms, like iOS and Android, are also considered; here the clients are in development. The devices that can connect to the server require a WIFI interface. The WIFI allows for a direct connection to the server and the corresponding screen. The user can interact with the content on the large display through the touch screen of the mobile device. The specific interactions that the smart devices support can be grouped into:

User registration: rescuers, refugees or citizen that reach a hotspot with one of the DIPP systems installed need to initially register and be included in the databank. The registration can take place as an individual or as a group. Furthermore, the system needs to capture in a central databank information like identity, type of users (medical rescuers, refugees, wounded, army), status values (e.g. level of health, equipment at disposition), possibilities of contact (e.g. radio frequency for rescuers with handheld transceivers). Once this information is gathered, the server stores it locally and, if a Internet connection is still available, it synchronizes with the other DIPP hotspots.

Displaying and downloading information: Once registered, users can manipulate the large display in order to highlight particular pieces of information, e.g. where is the nearest rescue shelter. At the same time, useful information can be downloaded to the mobile devices, like: maps with safe area, dangerous area, shelters, sources of food and water, temporary medical facilities, rescue

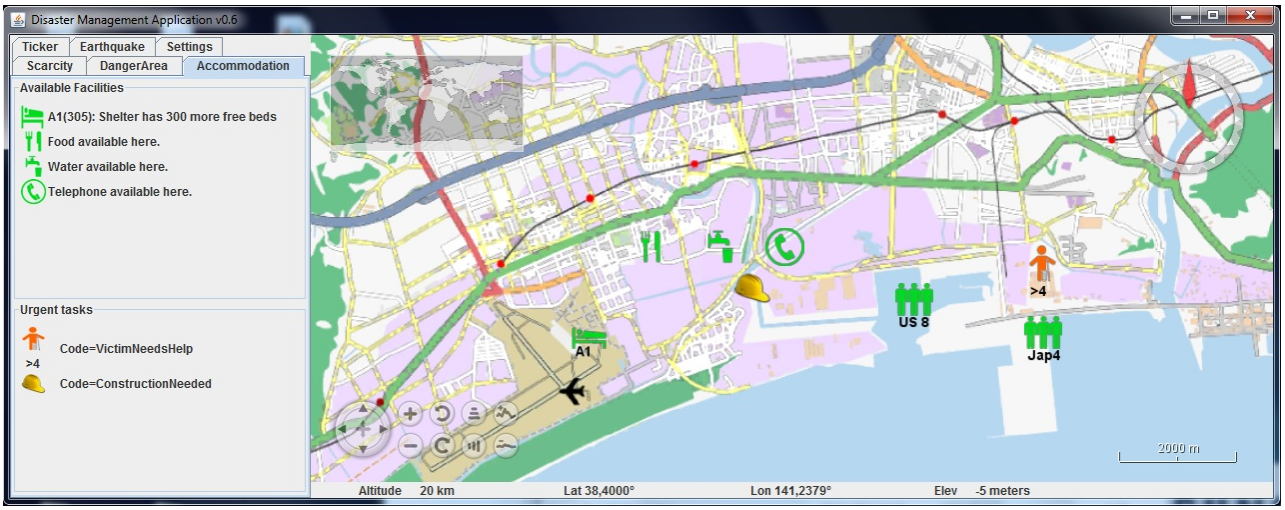


Fig. 6: DIPP application with accommodation layer enabled.

headquarters, current estimated position of rescue teams, areas already inspected or evacuated by rescue personnel, etc.; lists of teams, list of missing people, lists of people in a particular shelter, etc.; various media files, like videos and pictures that could give assistance to the rescue efforts. Depending on the type of the user or group, some of the mentioned operations are emphasized, both on the large display and the mobile system, in order to give a better overview of the existing possibilities. While this is primarily meant to speed up visual analysis on the map, it also has the advantage of only suggesting dangerous operations for the individuals with proper training (e.g. a citizen should not be given information on dangerous areas with missing people, as he might decide to investigate and help himself, which can be very dangerous without proper assistance).

Uploading information: All users have the possibility of uploading textual or media information. For example, a user can upload a video of a damaged building in the category damaged structures, and at the same time, give the coordinates on the map where this video was recorded. Therefore, it is important that the users first have to select the type of the information and its relevance before committing it to the system. If a user with limited authority posts certain data / information (e.g. a citizen that just registered), the information is labeled as *unverified* until it is validated by an authorized party (e.g. rescuers). Such information can be further enforced or denied by other users.

4.4 Agricultural scenario: iGreen

To demonstrate the flexibility and portability of the DIPP approach, its applicability has been tested in a second scenario in the agricultural domain. The iGreen project designs and realizes a network for knowledge and location based services in order to combine and utilize distributed and heterogeneous information sources of both public and private origin. The iGreen platform supports and optimizes energy-efficient, economic, environmentally friendly and frequently collaboratively organized production processes in the agricultural sector. iGreen offers the end-user standardized connectivity with intelligent technologies that provide collaboratively organized services based on actual and site-specific data. In this context, the large public display, where basically all important information is displayed, is mainly intended to be used by agricultural service supply agencies. This user group needs a fast, reliable and interactive overview on the overall situation (weather condition, positions of agricultural engines, seedtime, etc.). Their main goal usually is the optimization of cost and time efficiency. On the other hand, the farmers need to work with this data too. However, they are not interested in the global data, but in the localized information. Their mobile clients connect to the DIPP server in order to update and display the related information (maps, schedule, machine settings, warnings, etc.).

5. Conclusion and Outlook

Our DIPP approach makes a contribution to the problem of information distribution and information availability in case of natural disasters. When a country is hit by such a damaging event, local emergency personnel are dealing with coordination of local first responders, so there are no capacities for supporting of arriving supranational emergency personnel. This leads to a lack of information flow, thus hindering efficient and effective aid. The DIPP emergency management system is based on a client/server architecture and mobile device interaction over a WIFI network. It is easy to setup and use, which was the major goal of this system: not limiting the group of users due to cumbersome usability. In this way we take account of the diversity of

emergency personnel of various branches (e.g. medical personnel, engineers, fire fighters, police, military, and search & rescue parties). The approach of using public displays is suitable for a disaster and crisis management system. Unlike other systems specially tailored for expert users, familiar with the use of GIS and more complex software, our system is tailored for a wider, more diverse group of users. It is necessary to provide the variety of first responders with the needed and important information, without the need of a long training period or reading a user's guide on how to use the disaster and crisis management system. In order to verify the suitability of our proposed system, we conducted an informal evaluation with non-expert users, representing the variety and diversity of future users. The remarks we received after the cognitive walkthrough have been mainly positive, especially regarding the perception of the displayed information and the easy way of accessing it. The icons used for visualization were self-explanatory and comprehensible, aiding users to minimize the training period. However, one has to keep in mind that our system is still work-in-progress, some functionality still has to be added, but these issues will be addressed in future. The overall reception of the prototype system has been positive. For future, we want to extend the functionality of both server and client and also conduct a formal evaluation of the system, by performing a user study.

6. Acknowledgements

This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) as part of the International Graduate School (International Research Training Group, IRTG 1131) in Kaiserslautern, Germany. The work with regard to the iGreen demonstration scenario was funded by the Federal Ministry of Education and Research, Germany (BMBF, 01lA08005G).

References

[1] F. H. Abed, H. Zhang, and H. Zhang. Open source web-based GIS and database tools for emergency response. *In Proceedings of IEEE International Conference on Automation and Logistics (ICAL 2008)*. Qingdao, China, 2008.

[2] Bundesamt für Bevölkerungsschutz und Katastrophenhilfe (Editor). Katastrophenmedizin - Ein Leitfaden für die ärztliche Versorgung im Katastrophenfall. Bonn, Germany, 2010.

[3] A. Cheok, A. Mustafa, O. Fernando, A. K. Barthoff, J. P. Wijesena, and N. Tosa. BlogWall: displaying artistic and poetic messages on public displays via SMS. *In Proceedings of the 9th International Conference on Human computer interaction with mobile devices and services (MobileHCI '07)*. Singapore, Singapore, 2007.

[4] T. Chen, H. Yuan, R. Yang, and J. Chen. Integration of GIS and Computational Models for Emergency Management. *In Proceedings of 2008 International Conference on Intelligent Computation Technology and Automation (ICICTA)*, pp. 255–258. Hunan, China, 2008.

[5] E. F. Churchill, L. Nelson, L. Denoue, J. Helfman, and P. Murphy. Sharing multimedia content with interactive public displays: a case study. *In Proceedings of the 5th Conference on Designing interactive systems: processes, practices, methods, and techniques (DIS '04)*, pp. 7–16. Cambridge, MA, USA, 2004.

[6] P. Currion, C. Silva, and B. Van de Walle. Open source software for disaster management. *In Commun. ACM*, Volume 50, Issue 3, pp. 61–65. New York, NY, USA, 2007.

[7] H. W. Fischer. The role of the new information technologies in emergency mitigation, planning, response and recovery. *In Disaster Prevention and Management*, Volume 7, Issue 1, pp. 28–37, 1998.

[8] S. Greenberg and M. Rounding. The notification collage: posting information to public and personal displays. *In Proceedings of the SIGCHI conference on Human factors in computing systems (CHI '10)*, pp. 514–521. Seattle, Washington, USA, 2001.

[9] K. Gronbak, A. Rohde, B. Sundararjah, and S. Bech-Petersen. InfoGallery: informative art services for physical library spaces. *In Proceedings of the 6th ACM/IEEE-CS joint Conference on Digital libraries (JCDL '06)*, pp. 21–30. Chapel Hill, NC, USA, 2006.

[10] S. Hosio, M. Jurmu, H. Kukka, J. Riekki, and T. Ojala. Supporting distributed private and public user interfaces in urban environments. *In Proceedings of the 11th Workshop on Mobile Computing Systems and Applications (HotMobile '10)*, pp. 25–30. Annapolis, Maryland, USA, 2010.

[11] A. Huang, K. Pulli, and L. Rudolph. Kimono: kiosk-mobile phone knowledge sharing system. *In Proceedings of the 4th International Conference on Mobile and ubiquitous multimedia (MUM '05)*, pp. 142–149. Christchurch, New Zealand, 2005.

[12] K. Iizuka, Y. Iizuka, and K. Yoshida. A real-time disaster situation mapping system for university campuses. *In Proceedings of the 4th International Conference on Online communities and social computing (OCSC '11)*, pp. 40–49. Orlando, FL, USA, 2011.

[13] N. Kapucu. Interagency Communication Networks during Emergencies. *In The American review of Public Administration*, Volume 36, Number 2, pp. 207–225, 2006.

[14] J. Landgren and U. Nulden. A study of emergency response work: patterns of mobile phone interaction. *In Proceedings of the SIGCHI Conference on Human factors in computing systems (CHI '07)*, pp. 1323–1332. San Jose, CA, USA, 2007.

[15] T. K. Lui, Y. W. Huang, and J. Y. Chung. Interactive Wireless Electronic Billboard. *In Proceedings of the 2004 IEEE International Conference on Networking, Sensing and Control*, pp. 553–558. NY, USA, 2004.

[16] F. Marincioni. Information technologies and the sharing of disaster knowledge: the critical role of professional culture. *In Disasters*, Volume 31, pp. 459–476, 2007.

[17] M. Mes. Public safety in the aftermath of earthquakes. *In GEOconnexionUK*, Volume April/May, 2010.

[18] J. I. Na., N. Okada, and F. Liping. A Participatory Workshop Approach to Scenario Development for Disaster Relief, Response and Recovery Processes. *In Proceedings of the 2010 International Conference on Systems Man and Cybernetics (SMC)*, pp. 3433–3438. Istanbul, Turkey, 2010.

[19] L. Palen, K. M. Anderson, G. Mark, J. Martin, D. Sicker, M. Palmer, and D. Grunwald. A vision for technology-mediated support for public participation and assistance in mass emergencies and disasters. *In Proceedings of the 2010 ACM-BCS Visions of Computer Science Conference (ACM-BCS '10)*, pp. 8.1–8.12. Edinburgh, UK, 2010.

[20] E. L. Quarantelli. Problematical aspects of the information/communication revolution for disaster planning and research: ten non-technical issues and questions. *In Disaster Prevention and Management*, Volume 6, Number 2, pp. 94–106, 1997.

[21] I. Rauschert, P. Agrawal, R. Sharma, S. Fuhrmann, I. Brewer, and A. MacEachren. Designing a human-centered, multimodal GIS interface to support emergency management. *In Proceedings of the 10th ACM International Symposium on Advances in Geographic Information Systems (GIS '02)*, pp. 119–124. Mc Lean, Virginia, USA, 2002.

[22] R. Stephenson and P. S. Anderson. Disasters and the information technology revolution. *In Disasters*, Volume 21, Number 4, pp. 305–334, 1997.

[23] S. Thelen, D. Cernea, P. S. Olech, A. Kerren, and A. Ebert. D.I.P. - A Digital Interactive Pinboard with Support for Smart Device Interaction. *In Proceedings of the IASTED International Conference on Portable Lifestyle Devices (PLD '10)*. Marina Del Rey, USA, 2010.

[24] S. Underwood. Improving Disaster Management. *In Communications of the ACM*, Volume 53, Number 2, pp. 18–22, 2010.

Emotional knowledge engineering: Children as our innocent opponents in urban spaces' ownership

Anahita Mohammadi[1], **Mohammadmehdi Khabiri**[2]
[1]Architecture Department, Islamic Azad University of Beiza, Iran
[2] Architecture Department, Islamic Azad University of Khormooj, Iran

Abstract - *Children in our urban spaces are considered as elements that should be protected very carefully. The strange and unsecure physical and mental structure of these urban spaces changes parents from "parents" to constant guardians for children when they are in urban spaces. This inquiry aims at finding a method for clarifying the urban spaces and transferring the hidden concepts by simplifying them and finding communicational mediums. Vision is considered to be to realization and understanding color while roles as the conveyer element of data and information. A simple kind of relevance is explained which is formed on the basis of children's perception and intuition in relation with color. A new kind of classification and representation of conception in the emotional information engineering will be discussed here.*

Keywords: Children,Color,Urban Spaces,Vision,Invation

1 Introduction

Urban spaces are occupied by "us", the adults. All the functions of these spaces are totally regulated for us, and about children!?, ….. and children are transferred from one ward to another in a multi-fragment prison just like handbags in their parents' hands.- from kinder gardens to parks or playing rooms that undoubtedly all of them are totally protected and isolated from outside world- with this difference , that this time the interest and love toward them results in depriving them from freedom in urban spaces. These spaces are meaningless for children. In other words they spend their childhood in our meaningless urban spaces in dark.

2 Children as "aliens" in our urban spaces

In response to humans' need for developing a close and face to face relationship among people, our concepts of our urban spaces have reached to the highest level of quality. These needs have been expressed in varied forms and features from the beginning of the formation of human societies and we, as both producers of these needs and as the responsibles of answering them have coped with recognition and also removal of them in the framework of urban spaces with our own quality. All of us as human beings assume using these urban spaces as our natural and obvious right. On the other hand these above mentioned spaces have adjusted themselves with our innate diversity in the formation and renovation time. Although the above mentioned sentences seem to be nice and tangible to some extent, they just draw out a part of the reality and that reality also is interpreted and distorted by "us".

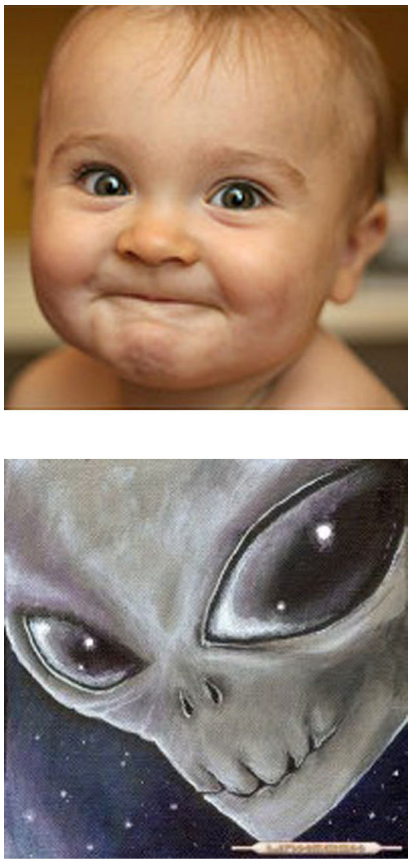


Figure 1 – Children as “Aliens”

Here, we, as dictators who interpret and present history and reality, have performed our roles well in hiding the truth. The fact is that in designing and renovating urban spaces' life, we have come across with a different kind of prejudice and classification. A kind of racialism that color and race was no more it's determining factors, but "age" is introduced as the most fundamental factor in this classification. We have accorded age as a criterion for being a human. In designing urban spaces children are not considered as a part of the owners of urban space but as the troublemakers that a sanctum must be defined for them. Scales, meanings, complicated relations and multi-dimentional functions of urban spaces are the reasons that we need a suitable medium for transferring information to the user and also a rather high level of understanding for the possibility of analyzing this bulk of information. This happens while in order to provide a possibility of suitable communication of children with urban spaces, their cognitive level should be increased. This process needs a time as long as their very childhood time. Since what is to be more emphasized is the recognition and definition of a proper crosscut communication tool among that set of basic and essential meanings by children. Here, like scientific/vagarous movies, we look for a kind of conceptual communication with aliens that unlike the general subject of these movies, their right of governing on the earth is actually reserved and undoubted.

3 Vision as the basic tool of reciprocal communication

Vision sense is introduced as the most evoluted entrance of the information in children. The evolution speed of this sense from the beginning of birth is like that we have to interpret it in the case of a mutation that occurs in short time intervals, in a way that the child's reaction from winking to the camera's flash at the birth time changes to following the moving objects or people in the 18 weeks, seeking the toys that are dropped from his hand in 32 weeks, using and orchestrating the eyes and hands in physical activities in 12 months, visual recognition in 20 months, naming symbols on the paper in 40 months, coloring inside the lines in forth years and reaching the visual alertness and being watchful of the environment in fifth years.

Figure 2 – Aliens and us (close encounters of the third kind)

On the other hand the statistics of more than 98 percentage of vision health at the birth time of children, changes this sense to a generalizable and essential tool for performing the relational medium role and reciprocal transformation of information among adults and aliens.

4 Color as the conceptual medium of transmission of potential information hidden in the space

4.1 Children and color

Children from the beginning are capable of exact recognition and segregation of colors without needing a lingual equivalence for defining and classifying them. The ability to recognize and segregate the colors from a number of months by children seems very rudimentary, while we comprise it with the complicate and comprehensive psychological topics based on colorful footprints remaining of them.

In another definition till children are child, they conclusively choose fantasy and color in the selection dilemma between reality and fantasy and also color and form. Reality and form are incorporated in children's world when their childhood is actually passed and they submit themselves to adult's world by the compulsion of age.

For children this is the color that is pure and precise and capable of showing the true meaning of what they have in mind, not the use of form and images.

The simplicity and immediacy of color and direct recognition and understanding of it by children, introduces

it as the most appropriate communicational language between us and these little aliens.

In effect the children's relationship with colors is a reciprocal relationship. On the one hand, children express their concepts easily by means of colors, while On the other hand colors contain a clear conceptual load of data for children.

While warm colors are full of energy and invitation for children, cold colors provide a neutral context for them.

Yellow and orange colors transfer a positive and pure thrill to the children. The red color has a special if it used prepared and appropriated for children. The colors of green and blue colors are neutral and without any powerful sensation for them.

While brown and gray colors not only show the negative sensational load for children but also they easily can transfer such kinds of sensational feelings to the children and cause that in children.

Observation and analyzing the brain activities in reaction to seeing colors explain this fact that combination of special colors beside the aforementioned colors can transfer clear messages to the brain. For example the combination of yellow and black directly dramatizes danger in children's brain.

Another advantage that introduces color as the most appropriate conceptual medium for transferring information. Is the existence of a specialized field in children's psychology which is shaped on the basis of colorful remains of them and has an antiquity of a number of decades. The "color" under discussion has become a familiar tool for both of the "types".

4.2 Color as a fundamental element in environmental graphics

There is a long time since environmental graphic has opened its place in our urban environments as an international language. Signs and colors are the most essential correlational factor between concepts in man's surrounding and the man himself. Colors in night and day shall introduce the spaces and make them recognizable for us. Their role exceeds from the impressive elements in environmental beautifying and at the same time transforms into the vital elements in environmental resurrection. Spatial legibility and resurrection of urban elements changes into such a vital mission that designers of environmental graphic inevitably have to link their professional activity field with the observations regarding psychology of colors. Because of the sensitivity of the subject, the design of the environmental graphic must have complete domination over investigational studies regarding color and culture, psychology of colors and spatial realization of colors to come to effective results in usage of color and form and dimensions in urban environment. The irony of the subject is that experts in the field of psychology of colors use children as one of their essential elements in researches related to psychology of colors because of the fact that children directly use their basic senses and intuition in understanding and being involved in psychological issues related to color, without using their little amount of knowledge. They participate in quality of what "we" have deprived them from their natural right by ignoring their presence among people and their right of cooperation in urban environments.

5 New interpreter layer as the medium of conceptual connection between two species

If we consider the concepts that are transferred by urban environments, as an informational layer, a new interpreter layer must be place upon the urban environment and become mixed with it. A colorful layer that translates the potential concepts of the urban environment to the puerile language and transfers to them in an understandable way. The new layer is full of pure colors. In fact this layer is not anything but pure colors and nothing else must that be. Using any formal metaphor changes this new layer to an insecure existence one more time that no more is responsible for its creator's aims. In fact any meaningful form used in this layer degrades it to the surface of the first layer that belongs to us "adults" one more time. In this new layer meaning should give its place to the sensation, perception and intuition because of the fact that these are children's communicational doorways with their environment.

The new layer must proceed in redefining urban concepts and hidden potentials and also changing them to decodable codes and signs by children. Positive potentials like places of aggregation, secure places, places full of energy, friendly place and so on or negative potentials like privacy places, insecure places, boring places, unfriendly places are defined in this colorful and puerile layer.

Children's interpreter layer, besides being simple and colorful, contains complicated concepts and multiple hierarchies and sub layers because the defined childhood in these topics has shaped on the basis of two essential viewpoints.

5.1 First point of view, psychological

First point of view, which is on the basis of children's perception and recognition level is according to the researches arose from children's psychology of colors. The concepts and feelings that colors represent for children, will face evolution and variation with their growing up, and in more aggressive phases even are segregated in the sexual form.

Figure 3 – "Impression sunrise"

Figure 4 –"Monaliza"

5.2 Second point of view, physiological

Second point of view is based on the improvement ability of visual sense that as it was mentioned before, in childhood it has a mutation-like progression. This level of progression shows its difference especially in the definition of visible distance and also the distance of children's attention and focus or in other words the defined depth of scope to them.

The mixture of these two progressions in children and simultaneously the consideration of these two impressive elements, naturally leads the children's interpreter layer despite its simplicity of perception to the conceptual complexity and multi-layer. This interpreter layer in fact is not unlike the works of Clod Moneh, the French painter of the 19th century that has been drawn in the Impressionism style, or the smile in Monalisa- the famous work of Davinchi, the Italian painter of the 16th century. Because of the fact that the colorful layer not only is interpreted according to the perception of the addressee to different form, but also the location and his mental conditions are effective in the form of impression and perception which is due to it.

In a way , just like Moneh's paintings, pure colors sit next to each other and by keeping their original identity, with a different distance that the viewer will have to the piece of art, colors will optically be mixed and will morph to different colors with different concepts.

6 Case in point

Naturally besides legibility of environment for children, their interpreter layer must be able to transfer parent's intended concepts to them. For example access network is interpreted to the danger's confines by means of colors and maybe brown and grey become dominant in this confine. Squares and parks change to interesting places and it's expected that they're represented in red, yellow and orange colors. This is while according to present viewpoints maybe high buildings would represent themselves as backgrounds with natural colors like blue and green. Here the existence of a kind of mathematical algorithm as the direction axis of effective and purposeful specification and implementation of this colorful layer seems to be necessary that itself is the opening way for new topics in the continuance of present topic.

Figure 5 – Present Urban Block

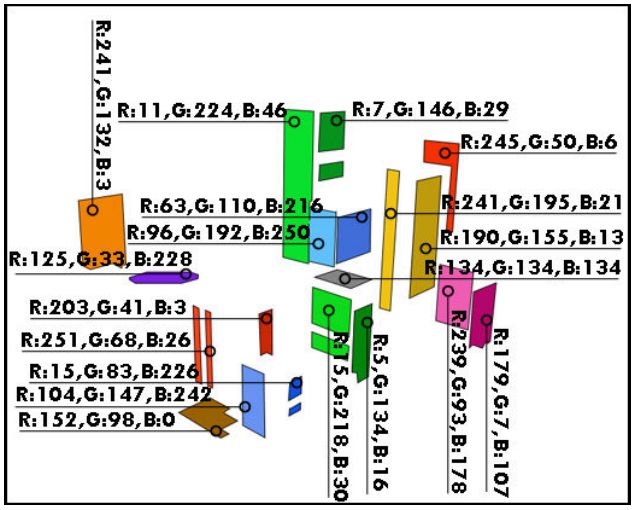


Figure 6 –"Kids' Interpreter Layer"

And finally this colorful layer should not be considered to be constant and permanent, but it should be able to mutate with the environmental mutation and necessities and also to converse to a new sustainable environment.

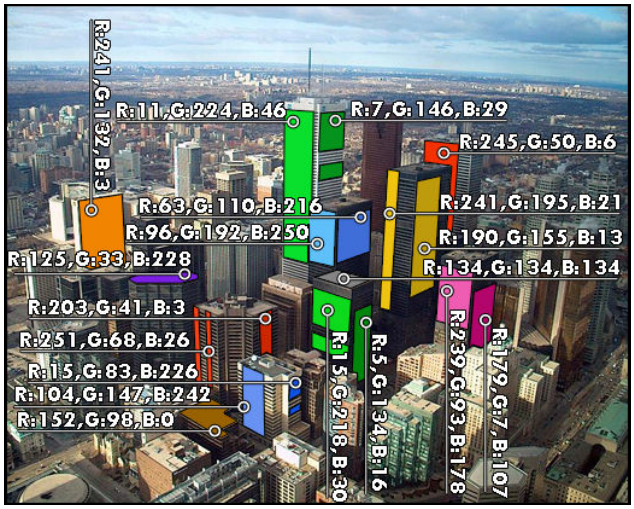


Figure 7 – Final result of the colorful "Invasion"

7 Conclusion and Extension

What the topics of this article introduce is the effort for opening a new field of relation between two different species of "us" and children for reaching a permanent and constant peace and collective and square use of available sources, meaning our urban environments for accessing to the utilize methods for realization of this desire in a thoughtful way among psychologists, artists, doctors, designers of urban environments and also mathematicians for conversing patterns governing on children's interpreter layer to applicable distances it's needed to reach to a proper understanding of these intellectual and colorful layers before their implementation. And if here , our first step was meant to transfer our intentions to children , sure the next one should involve children in the process so they would be able to transfer they needs and intentions to "us".

8 References

[1] Abcrcrobie, C. F. Stanley , 1991, places for play , urban , open space Academy editions , London.

[2] Broto , carles , 2006 , Great kid's spaces , translated by amber ockrassa , links.

[3] Church , Josef , Steven Gorge , A psychology of the grewing person .

[4] Azimi , Sirous ,2002 , child's psychiatry , Amir Kabir .

[5] Craw , Alice W. , 1998 , child's psychiatry , translated by Hamedani , Moshfegh , Amir Kabir .

[6] Alkaiid , David , 1996 , Child's growing and education in Piazheh philosophy , translated by Naely , Hosein , Astan-eh-Ghods .

[7] Dr. Amably , T. , 1997 , Kids' creativity , translated by Dr. Ghasem-zadeh , Hasan , Azimi , Parvin , Donyayeh now , Tehran.

[8] Izadpanah Jahromi , Ayda , 2005 , Kids : city & play – process , basis and regulations of kids' playground design - , National municipalities organization .

Self-optimizing Production Program Planning for Product Ramp-ups

A. Günther Schuh[1], **and B. Maik Schürmeyer**[1]
[1]Institute for Industrial Management (FIR), RWTH Aachen, Germany

Abstract - *Because of decreasing product life cycles and increasing product diversity, producing companies are confronted with a growing number of product ramp-ups. The associated high effort in the production planning and control (PPC) for ramp-up products can only insufficiently be reduced using existing planning algorithms, because of a lack of reliable historical information. This problem is highly significant for the production program planning (PPP): as the first step of PPC, its lack of information is exceptionally high. The goal of current research activities is to establish a concept for reliable and fast PPP for companies with numerous ramp-ups. To reach this goal, a model for a cybernetic PPP is currently developed. For this interdisciplinary and cybernetic design of the PPP, the Viable System Model (VSM) is used as frame of reference. It enables a transparent and efficient handling of highly complex and interdisciplinary structures, processes and information flow.*

Keywords: Information Management, Cybernetics, Production Planning and Control

1 Introduction

The last decades showed a clearly trend to rising product variety as well as to steadily decreasing product life cycles [1, 2, 3]. This results in a drastic increase of production ramp-ups.

According to WACK, the average number of ramp-ups at car manufacturer Mercedes-Benz Cars has increased more than three times a year for the last two decades [4]. A study, in which 225 ramp-up situations of 100 car manufacturers and suppliers were inspected, showed that the serial production ramp-up is still burdened with major problems. In 33% of cases the economic goal could not be achieved and in 50% of them the technical goal was missing [5, 6].

1.1 Production Planning and Control

One of the reasons behind the poor performance in ramp-up projects are the problems in production planning and controlling (PPC). The PPC for ramp-up products are particularly challenging, because the information needed for planning (e.g. product structure, sales projection etc.) is either missing or is not of sufficient quality in terms of actuality, accurateness or granularity.

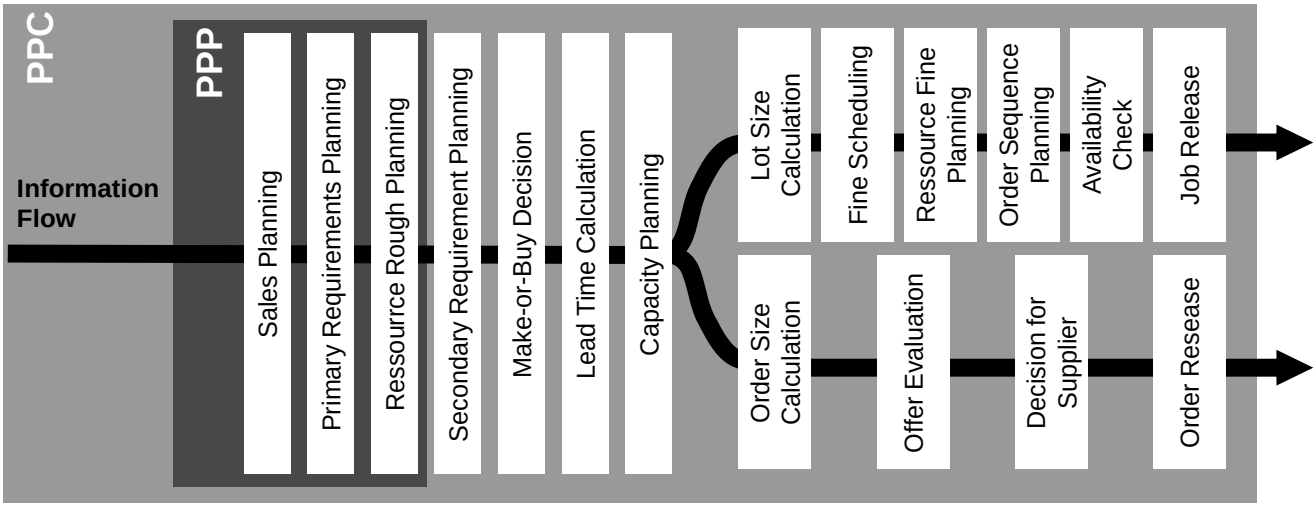


Figure 1: PPC process for make-to-stock production [8]

Reasons for this are the dynamic environment changes on the one hand and changes in the production process in this phase of product life cycle on the other hand. Furthermore, the underlying target system and other restrictions of PPC in serial production can vary from those in stable serial production. For example, the ramp-up restrictions could be the uncertain effectiveness of resources, the limited purchasing volume or the quality of externally procured greige [7].

1.2 Production Program Planning

The production program planning (PPP) is the initial step of PPC, see Figure 1 [8]. Therefore, PPP plays a specific role for the entire PPC process. STICH emphasizes this particular role of PPP as follows: "The different conditions in the ramp-up phase compared to serial production are exceptionally affecting the production program planning." [7]

Since PPP as the initial step of PPC does not have any input information from preceding PPC-processes, there is a higher lack of information compared to other PPC-processes. The following list shows the potential information deficit in PPP regarding the information availability and quality (actuality, correctness, and granularity).

- Sales history, composition of product characteristics
- Evaluation of forecasting method and its parameterization
- Practical inquiry and orders, reserved in- and outflows, internal demands
- Capacity and error rates of resources (personnel, assets, utilities)
- Bill of Material, work plans, different cost rates

These information deficits can have a negative impact on the quality of the production program. Possible effects of insufficient information availability or quality are:

- Faultiness of production program
- Late completion of the production program
- Limitation of production program to predictable products

According to the law of error propagation, these faults, delays and limitations may partially increase in the following steps of the PPC process [9] . As a result of planning insecurity and error propagation, the PPC is confronted with several risks:

- Overproduction or underproduction
- Loss of market shares by unmet customer requirements
- Faulty purchase quantity in general agreement with suppliers

These risks may involve a concrete financial loss. Overproduction affects the capital commitment costs, obsolescence expenses and inventory costs. In case of underproduction, the opportunity costs for unused production resources occur. Loss of market shares can lead to loss in revenue. Too high purchase quantity in general agreement with suppliers may lead to high carrying costs for capital commitment, obsolescence and warehousing. Too low purchase quantity leads to production downtimes and therefore to possible contract penalties.

To suppress the low quality of planning, particularly the high-wage countries are incorporating extremely sophisticated and time-demanding planning methods [10]. Higher planning complexity induces not only higher personnel costs but also brings about the risk of time setback. Especially in the case of a new product, the "time-to-market" can influence the success of an innovation. These conflicting goals are the first main problem of PPP at ramp-ups and will be depicted as "Dichotomy Stability vs. Speed".

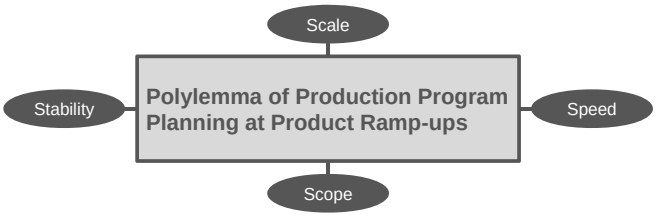


Figure 2: Polylemma of Production Program Planning at Product Ramp-ups

It is possible to reduce the "Dichotomy Stability vs. Speed" by limiting the production program. With such a limitation of production and vending amount of goods, the planning quality can be improved by specialization and its accompanied economies of scale. However, if the production program is kept low because of planning, it may be that certain customer requirements cannot be met which in turn may influence the company's success negatively. This conflicting goal is the second main problem of PPP in ramp-up situations and it will be depicted as "Dichotomy Scale vs. Scope". In the following, both of these goal conflicts will be depicted as the "Polylemma of Production Program Planning at Product Ramp-ups", see Figure 2.

2 Objective

The objective of this research is to develop a methodology to reduce the "Polylemma of Production Program Planning at Product Ramp-ups", which consists of the following dichotomized dimensions:

- Stability: High planning quality should warrant a stable production program.
- Speed: Low planning complexity should anticipate the product-to-market introduction delays and high personnel costs.
- Scale: With the use of specialization and learning effects, the economies of scale should be used in planning.
- Scope: With a strong focus on customer needs, the diversification effects should be used.

To achieve this goal, a model that allows a self-optimizing PPP is being developed. In order for PPP to be self-optimizing, it must be able to learn from earlier planning cycles. It also has to be able to automatically recognize and compensate inner and outer faults. To be able to do so, high resolution information must be available in adequate granularity, correctness and actuality; for fault compensation, the system must be able to adapt itself to the new circumstances without much effort.

Since cybernetics fulfills these requirements completely, it will be used as the main principle for the creation of the self-optimizing PPP. The model "Cybernetic Management of Versatile Production Systems" by BROSZE investigates cybernetic forms of production systems intensively. It serves as the reference model for this research [11].

3 Conceptual approach

The conceptual approach is split into three solution segments

- Information demand analysis
- Sensitivity analysis for evaluation of information criticality
- Creation and design of a VSM-based cybernetic PPP

3.1 Information demand analysis

The aim of the information demand analysis is a detailed display of information-flow in the conventional PPP. For this purpose the main tasks of intermediate PPP are identified and broken down into several planning processes. For every identified step of the process, methods, formulas and algorithms needed for the solution will be composed. Based on these methods, formulas and algorithms, the needed (input-) information and the generated (output-) information will be passed to PPP and consolidated in the following steps.

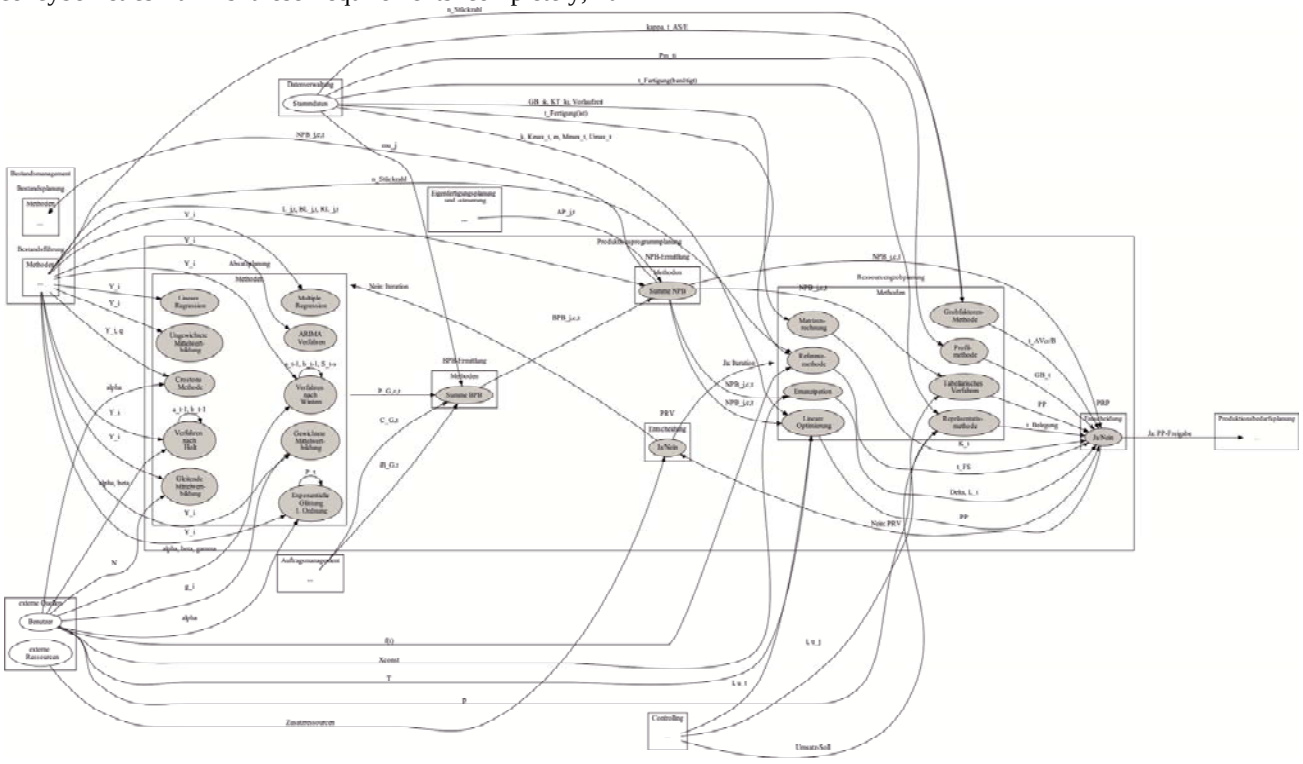


Figure 3: Information map of PPP based on Aachen PPC-Model [8]

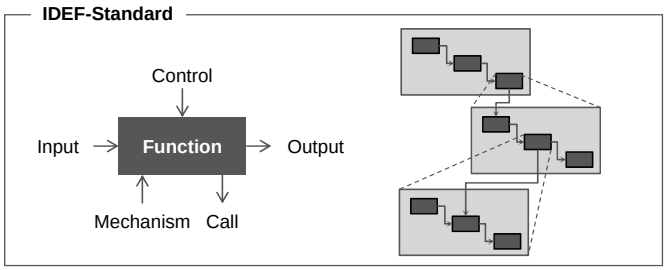


Figure 4: Modelling of the information flow by means of IDEF-standard [12]

The base of this research is the definition of boundaries of PPP system according to system theory. The result is a detailed map of the overall required information in PPP. To avoid unnecessary calculations and to warrant completeness of the information map, the analysis uses the pull-principle: depending on the aspired result (intermediate production program) the process will be backtracked to the predefined boundary of PPP or rather PPC.

The modeling steps are done by means of IDEF-Standard, cf. Figure 4. This systematic approach allows to generate a transparent model with high complexity.

One temporary result is shown in Figure 3. It shows an information map of PPP referring to Aachen PPS-model. Future research results will concentrate to drill down those information flows in terms of granularity. By that means a "high resolution PPP" is to be created.

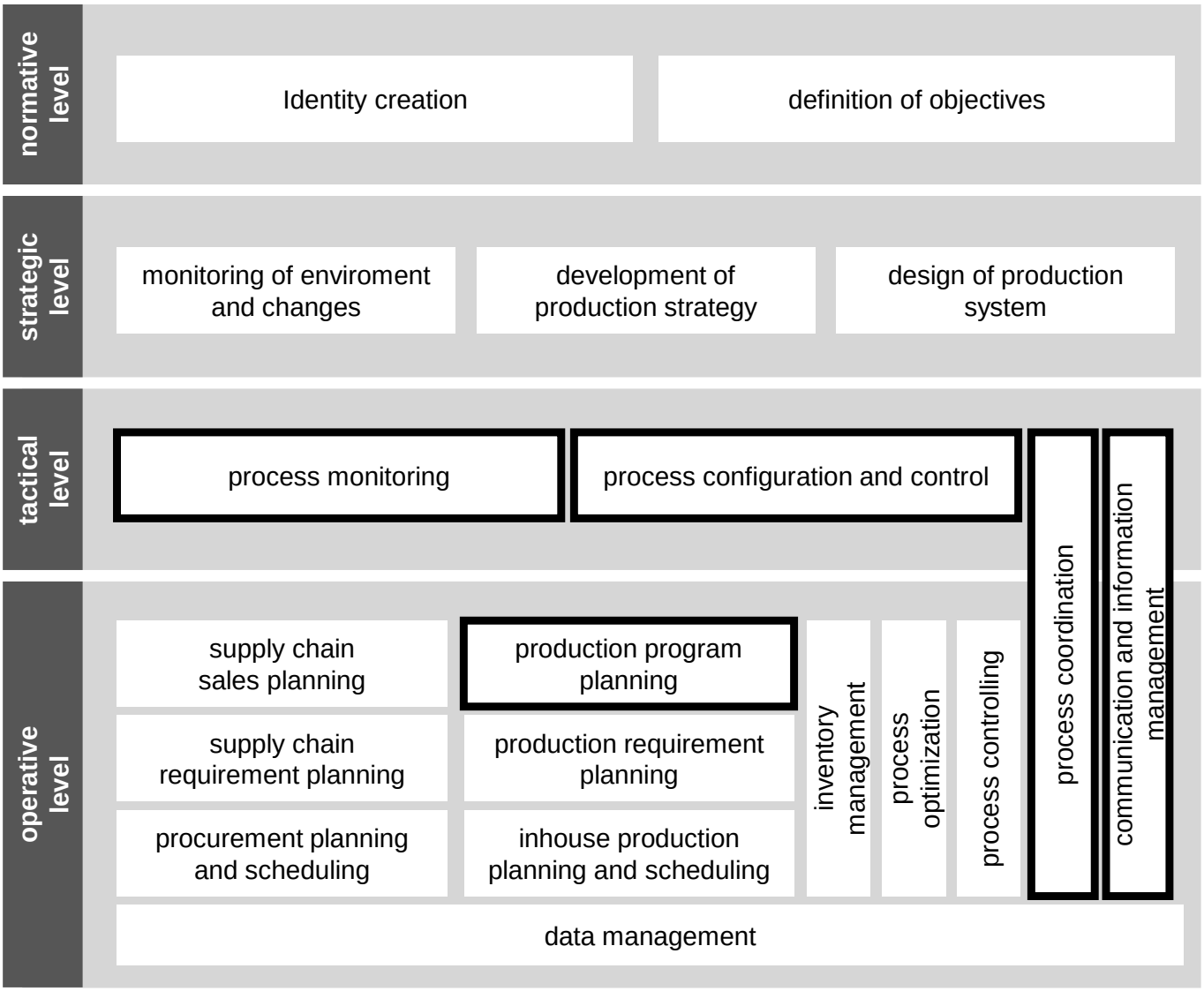


Figure 5: Extended model of PPC-tasks, cf. [11]

3.2 Sensivity analysis

The aim of sensitivity analysis is to identify the criticality of the input information. This evaluation is carried out in the following steps:

- Definition of a ramp-up specific target system for PPP
- Construction of a quantitative evaluation scheme for the quality of input information based on the predefined characteristics (availability, correctness, actuality and level of detail).
- Explanation of the correlation between the quality of respective input information and its effects on the target system and derivation of a sensitivity-index for the necessary input information.

3.3 VSM-based cybernetic PPP

Creation and design of a VSM-based cybernetic PPP aims on a PPP as a part of the Cybernetic Management Model for Versatile Production Systems according to BROSZE, see Figure 5 (The highlighted labels affect the resent research activities). For this purpose, the VSM-systems 1 to 5 of PPP and ramp-up management are defined and integrated into the reference model. Additionally, the necessary communication infrastructure comprised of process routing centers, process control centers and process coordination centers is formed.

Furthermore, the operational VSM-systems are identified as alternative sources of information and integrated in the existent VSM-structure. The possible information sources are for example the different entities of ramp-up management or a person as an individual part of the entire system. Special attention is turned on creation of knowledge base to warrant interdisciplinary communication and the use of learn effects.

4 Conclusions

The paper in hand describes a way to form a cybernetic production program planning for companies with numerous ramp-ups. The first step describes the complete information flow that warrants a high quality production program in terms of actuality, correctness and granularity. In the second step the sensitivity of necessary input information on the result of the PPP is analyzed in order to evaluate the information criticality. In the third step, the information flow is arranged in such a way that an automatic and qualitative high-grade planning is made possible. Additional information sources for particularly critical information are identified and integrated into the process of planning. By that means the planning quality in ramp-up situations is improved. Furthermore, the information flow is arranged according to principles of cybernetics so that the faults can be compensated in best possible way and learn effects can be used reasonably. The results of the explained research activities can allow the ramp-up intensive companies to create reliable production programs in short time. Learn effects can be used without neglecting diversification. In conjunction with cybernetic arrangement of other parts of production planning and control, the development of cybernetic-operational ERP-systems is strongly pressed ahead.

5 References

[1] Horst Wildemann. “Anlaufmanagement : Leitfaden zur Optimierung der Anlaufphase von Produkten, Anlagen und Dienstleistungen”. 8. Aufl. München: TCW Transfer-Centrum, (pg. 5), 2009.

[2] Andreas Romberg; Martin Haas; Dieter Hermenau. “Der Anlaufmanager : Effizient arbeiten mit Führungssystem und workflow - von der Produktidee bis zur Serie” (pg. 10 f.). Stuttgart, 2007.

[3] Sabina Fjällström: “The role of information in production ramp-up situations”, Chalmers University of Technology, Göteborg, 2007.

[4] Karl-Josef Wack; Thomas Bär: “Grenzen einer digitalen Absicherung des Produktionsanlaufs - Limitations of Digital Ramp-up Validation”. In: Zülch, Stock (Ed.): “Integrationsaspekte der Simulation” (pg. 45-52)- Karlsruher Institut für Technologie (KIT), Karlsruhe, 2010

[5] Frank Straube: “E-Logistik: Ganzheitliches Logistikmanagement”. Springer, Berlin, 2004.

[6] Frank Denzler: Modellanalyse von Lieferantenbeziehungen in Anlaufprozessen. 1. Aufl. München : TCW Transfer-Centrum (pg. 5), 2007.

[7] Christoph Stich: “Produktionsplanung in der Automobilindustrie – Optimierung des Ressourceneinsatzes im Serienanlauf”. Kölner Wissenschaftsverlag, Köln (pg. 15), 2007.

[8] Günther Schuh: “Produktionsplanung und -steuerung : Grundlagen, Gestaltung und Konzepte”. 3., völlig neu bearb. Berlin : Springer, 2006.

[9] Werner Meyer-Eppler, Werner: “Grundlagen und Anwendungen der Informationstheorie”. (pg. 41), Springer, Berlin, 1969.

[10] Christian Brecher: “Integrative Produktionstechnik für Hochlohnländer”. Springer, Heidelberg, 2011.

[11] Tobias Brosze: “Kybernetisches Management wandlungsfähiger Produktionssysteme”. Apprimus, Aachen, 2011.

[12] J. Mayer: “IDEF 1 – Information Modelling”, 1992.

Information Management for the Competitiveness and Competence of the Small and Medium-sized Enterprises

Jose-Melchor Medina[1], Isabel De la Garza[2], and Alberto Mora[3]
[1] Facultad de Comercio y Administración. Universidad Autónoma de Tamaulipas. Victoria, México
[2] Facultad de Comercio y Administración. Universidad Autónoma de Tamaulipas. Tampico, México
[3] Facultad de Comercio y Administración. Universidad Autónoma de Tamaulipas. Victoria, México

Abstract - *The literature contains contrasting positions regarding the relationship between information technology and competitiveness in enterprises. Some studies argue that information technology is not a source of competitiveness. However, there are also a considerable number of studies that show how technology has helped organizations improve areas such as innovation, productivity, efficiency, decision-making, customer satisfaction, etc. Such improvement is then translated into a gain in the organizations' levels of competitiveness, which in turn helps them face competition. This research analyzes the degree of influence that the use of information management has on the competitiveness and competence of small and medium-sized enterprises. The empirical study was carried out in Tamaulipas (Mexico) through regression analysis. The results show that the efficient use of information has an effect on the two dependent variables, but, with a higher level of influence on competence (leadership in innovation, and in monitoring competitors).*

Keywords: information, competitiveness, competence, SMEs, IT

1 Introduction

Technology and information have become important tools that have been used by organizations not only to succeed but also to survive in a globalized world. The argument that enterprises that use Information Management (IM) effectively are more likely to attain higher levels of organizational performance and competitiveness is certainly not new. As a result of this, many enterprises worldwide have started to allocate a great deal of resources for the generation and application of information and knowledge aiming at improving their performance and competitiveness. However, little empirical work has been done which contributes to our understanding of the extent to which Mexican enterprises make use of this valuable tool.

The purpose of this research is to analyze the relationship between IM and the performance of an organization. In particular, the research looks at the relationship between IM and levels of competitiveness and competence in small and medium-sized enterprises (SMEs) in the central part of Tamaulipas, Mexico. We believe that knowing about information management practices within organizations is crucial to the development of institutional policies aimed not only to survive, but also to obtain higher levels of competitiveness and competence at the regional, national and international levels.

2 Literature Review

2.1 Information Management

Most, if not all, enterprises depend on information technology (IT) for the accurate and prompt management of information. IM can be defined as the economic and efficient production, control, storage, retrieval and dissemination of information, in order to improve the performance of the organization [1]. However, there is a challenge that many organization face these days. Once big amounts of data have been collected from the entire organization a question is raised: Now what do we do with them? In the efforts of contributing to the organizational effectiveness, the impact of the information remains hidden until it is removed or lost [19]. That is to say, information makes sense only if someone uses it for something; therefore, the critical areas of information transfer and storage must be closely and accurately defined for each system used in the firm [1].

The quality principles advocated by Deming, Ishikawa, Juran, Crosby, etc. for products quality are also used in the improvement of the information, applied to the problems of production of quality data output, where each information product has an intrinsic value for the user. Information quality is defined as the measurement of information technology data output in terms of accuracy, opportunity, completeness, reliability, and relevance. However, according to Lillrank [13] the most widely used definition of information quality is given by the American Society for Quality and ISO 9000-2000; such definition is based on customer satisfaction and places a strong emphasis on the idea that the requirements should not only be met for their own sake.

The emergence of information as a productive factor and development engine is now becoming evident in the wider society. The IT potential for the improvement of information performance in the organization has been widely recognized, since the availability of reliable information sources is a key

component in the decision-making processes of the executives as users [12]. These sources are selected as they are thought to be useful and therefore will offer the highest quality of information; there is also evidence that they help in the improvement of performance indicators such as data accuracy, speed in decision-making processes, effectiveness and ability of data analysis. Therefore, there is a need to take into account the following ideas about IM:

- The conception of IM should consider a transition from the focus on the information process and storage to that centered on its use and share [4].
- IM should be focused on people as the essential aim and consider IT as an enabling factor, perhaps necessary in the effective and satisfactory use of information [15].

Undoubtedly information is an intrinsic component in almost all the activities in every organization to the degree of becoming transparent. This is because it is the means through which people express, represent, communicate and share their knowledge. Marchand et al. [15] highlight that it is the use of information which has an influence in the creation of the business value through four strategic priorities: (i) minimizing financial, commercial and operational risks, (ii) reducing the costs of transactions and processes, (iii) adding value to customers and markets, and (iv) creating new realities through innovation.

2.2 Competence

For the purposes of this study, competence refers to that situation in which two economic entities (enterprises) engage in a constant struggle to sell their products and services in the market. That is why organizations view IT as a tool for the gain of competitive advantage that can serve to thwart the competitors' force.

The impact of the IT investments on their performance in business and on IM has been researched [14]. However, it seems that firms have not taken full advantage of such research findings, at least in Mexico and in the region where this study took place. That is why the amounts of money spent by firms in technology continue to grow excessively. Unfortunately, such expenditures continue to result in little or no benefit for them. On the other hand and as an example, in the financial sector, Hauswald and Marquez [7] have argued that the information process points to the need for investments made in technology to be productive. Although, it is important to highlight the fact that the more accurate the information is the more costly it becomes [3]; IT, though, can help make more opportune decisions and maintain rapid communication in the competitive environment in which we are now immersed.

IT appears to be changing the competition structure, helping large organizations make their administration job easier [8]. These authors claim that IT helps manage large volumes of structured and unstructured transactions, and helps collect and share information beyond a country's boundaries. Similarly, technological processes have affected the production and availability of information. According to Dell'Aricca and Marquez [5], this has served to change the competition nature in the markets, as an enterprise which possesses rich data related to users is more likely to focus its efforts on increasing the levels of competition in the market, lowering prices, as well as thwarting the competitor's forces.

Diverse studies focus on aspects related to competition and their competitors [2008]. However, it is important to analyze such aspects from other perspectives such as Michael Porter's theory or Classic, Neoclassic and Austrian theories. Likewise, due to the market's flaws and the delay in information processing at some enterprises, it is thought that those enterprises which take information seriously can turn it into a highly profitable opportunity advantage [6], considering that competition increases the intensity of the demand for information, especially accounting information [9]. For example, banks acquire information in order to *soften* the competition's loans and expand their markets [7]. Therefore, a good IT infrastructure can become a greater competitive advantage for enterprises. According to Dewan and Mendelson [6], IT is costly, but it becomes twice as much costly for competitors if they are to continue to compete against those organizations with a good technological base already in place.

We now proceed to present the hypothesis of our work for this construct:

H_1. Information Management plays an influential role in helping the SMEs to face and thwart competition.

2.3 Competitiveness

Enterprise competitiveness means to achieve an equal or higher profit than that of the competitors in the market. According to Lavon and Todd [10] competitive advantage is a phenomenon that occurs when a firm experiences returns that are superior to those of its competition (rents). So, it is known that information and knowledge are two factors which have a remarkable impact on the conception and sustainability of the competitive advantages for the organizations. IM provides organizations with the opportunity to either activate their new competitive strategies or to detect their competitors' response as a way to restructure the industry. Nevertheless, the enterprise does not obtain any competitive advantage by merely having more computers at their disposal, but by being able to use them. More specifically, enterprises obtain competitive advantages by strategically applying the information generated or contained in them.

However, investments in IT may not have an immediate impact or add value to a firm and are, therefore, more likely to be reflected in future profit streams [11]. Mendelson [16] developed a metric which quantifies the ability of an organization to process information and make fast and

effective decisions in a highly dynamic environment. He coined it as *organizations' intelligence quotient* which is based on the principles of an organizational architecture focused on the effective information flow, the speed in decision-making processes and the utilization of the knowledge resources when the environment generates big amounts of data whose effective process is key to success.

Both organizational performance and competitiveness need to take into account both financial measures as well as operational performance measures (non-financial) such as market share, introduction of new products and services, product quality, marketing effectiveness, reputation improvement, flexibility, and operations promptness and productivity [15]. We need to place a strong emphasis on the organizational design that facilitates the vertical and horizontal information flows that aim to achieve the organization's objectives. In a similar vein, Melville et al. [17] define the term *Business Value of the IS/ICT* (Information Systems / Information Communications Technology) as the impact of the IT on organizational performance observed at the level of intermediate processes and overall organizational level, including an impact on efficiency and competitiveness.

The hypothesis for this construct:

H_2. Information Management plays an influential role in helping the SMEs achieve higher levels of competitiveness.

3 Method

Both technology and information now play a crucial role in the performance of organizations, and some scholars have adopted a hermeneutics-critical approach to examine their uses and implementation. Thus, SMEs that make use of the notion of IM were selected for this study. In particular, the relationship between IM and the enterprises' performance in terms of gaining competitiveness and facing competition was analyzed in this research.

The process followed to attain the aim started with a review of the state of the art on IM, competitiveness and competence. The operationalization of the variables was carried out as follows:

- Dependant variables: Competitiveness (financial performance, market share, innovation levels in products/services, customer satisfaction) and Competence (leadership in innovation, keeping track of competitors, competence information).
- Independant variable: Information Management (strategic use of information, participation of key staff members in information management, continuous acknowledgement of information processes).

The empirical work was carried out in the central region of the state of Tamaulipas, Mexico. To do so, a 5 point Likert scale questionnaire was designed and piloted with 12 enterprises. This pilot stage served to identify those items which did not have the minimum recommended statistical load. The composition of the final version of the instrument was as follows: Four items related to Competitiveness; three to Competence; and five to Information Management, in addition to the general data items.

According to the Mexican Entrepreneurship Information System's Website (http://www.siem.gob.mx), as of January 2012, there are 1224 SMEs in Tamaulipas (excluding enterprises with fewer than 10 employees). Of them, 161 belong to the geographical zone under investigation. Unfortunately, due to the poor research culture in this zone, the final sample was made up of 46 enterprises (two questionnaires were administered to each enterprise: 92 valid for their analysis). The respondents of the questionnaires were (1) either the general manager or the owner, as they both make use of information on a daily basis, and (2) the head of the informatics department. The respondents were given one week to answer the questionnaire so that they could complete it freely and with plenty of time. The analyzed businesses represent all types of enterprises in a transversal study. The data were analyzed through descriptive statistics and a regression analysis technique using SPSS software 18th version (PASW Statistics).

Results

The size of the participating enterprises is distributed as follows: 14% have between 11 and 20 employees; 22% of them have 21 to 30; 28% between 31 and 50; and 36% between 51 and 100 employees. Regarding the type of activities that the enterprises do, those that belong to the trade (commerce) sector dominate, accounting for 44% of them.

Table 1 shows the reliability degrees of each of the variables measured with the Cronbach's Alpha. In order for a variable to be considered acceptable, its value needs to be greater than 0.7 [18]. If so, it indicates that the questionnaire is valid; and therefore, its results can be interpreted as reflecting the current reality. The whole set obtained a value of 0.802.

Table 1. Cronbach's Alpha Results Distribution

Variable	Cronbach's Alpha
Competitiveness	0.701
Competence	0.752
Information Management	0.865

After the descriptive analysis, it is important to indicate that according to Chin [2]: R (Relation) represents the *path coefficients*, which should obtain a value of 0.2 if they are to be considered significant, with above 0.3 being an ideal value. R^2 on the other hand, indicates the variance explained by the variable within the model. This should be equal or greater than 0.1, as lower values provide little information even if they are significant.

After that, the regression analysis was conducted with the aim of showing the inferential data. Table 2 indicates the results

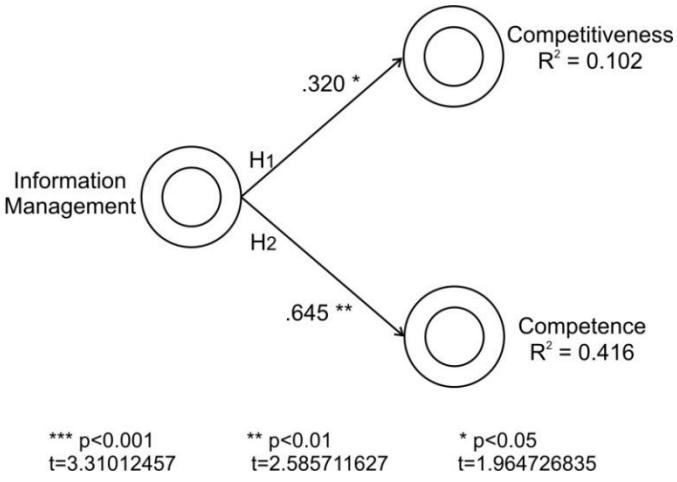
of the two proposed hypotheses. It shows the existing level of relation, the explained variance, as well as the obtained significance level, whose minimum value to accept a hypothesis needs to be equal or lower than 0.05 (at least 95% reliability).

Table 2. Hypotheses Assessment

Hypothesis	R	R^2	Sig.	Comment
H_1. IM →Competitiveness	.320 *	.102	.034	Accepted
H_2. IM →Competence	.645 **	.416	.002	Accepted

Figure 1 shows the assessed research model, which includes a graphical representation of the data as stated on tables above. It also includes the levels of relation between the independent and dependent variables with their respective hypothesis.

Figure 2. Assessed Research Model



The figure shows that the two proposed hypotheses were accepted. The analysis reveals a strong relation between Information Management and Competence, as they obtained a high level of relation (.645), a high level of explained variance (41.6%), with a reliability degree of 99%. This means that the efficient information management has served to outpass or thwart competence in terms of keeping a detailed record of internal and external data, especially data about the competitors' activities. Such a record of information has allowed the enterprises to develop more alternatives for their decision-making processes. In particular, their efficient information management process has enabled them to develop appropriate courses of action for their own benefit, as well as to defend their market position, which has been earned on the basis of hard work.

A similar issue seems to apply to Competitiveness. The obtained data out pass the minimum values as recommended by scholars. This seems to suggest that the appropriate use of information has assisted the SMEs in keeping an appropriate financial performance, in keeping or gaining a market niche, and in maintaining their customers' satisfaction.

4 Conclusions

The goal of this work is to determine the degree of influence that Information Management has on the competitiveness and competence of the SMEs. The proposed hypotheses have been tested; however, it is necessary to state that even though some of the studied enterprises have achieved an advanced stage in the management and use of information, none of them has validated empirically these ideas or designed an effectiveness measurement to determine if an enterprise is appropriately managing and using information.

Three main contributions to knowledge can be derived from the results obtained in this research: *i*) the SMEs make use of information without a methodology or in a systemic way; they seem to be simply reacting to changes in the market or in the competitors' strategies, *ii*) IM is basically the fact of having precise information for decision-making purposes. They seem to view quality information as an asset which has enabled them to use it moderately; and in a way they have managed to reduce costs. This has been reflected in their leadership in products/services and in monitoring competitors; and *iii*) regarding competitiveness, with their strategic use of information, these SMEs have been able to obtain financial stability, to generate innovations in products/services, and in processes. More importantly, however, is the fact that with their strategic use of information, the SMEs have been able to maintain their customers' satisfaction, which allows them to generate a virtuous circle in their best interest.

However, the dynamic nature of technology continues to raise questions that can be good subjects for future research projects. In addition, it is strongly recommended that the SMEs continue investing in technology and staff training and development so as to improve their profesional performance. Furthermore, the findings of this research suggest that little progress has been made in terms of information management, especially in competitiveness. Therefore, more knowledge in this respect still needs to be generated that can be useful for the different stakeholders in the knowledge society.

These findings give us an outlook of the current situation in Mexico. While they cannot be generalized to the whole country, they certainly enable us to realize that the concept of knowledge management is still not taken seriously at least in the region in which this study took place. The evidence presented here suggests that the Mexican SMEs are still not able to use the information generated within them properly. Therefore, we suggest that this is topic that needs more attention, and certainly more research.

5 References

[1] Best, D.P. (2010). The Future of Information Management. *Records Management Journal*, 20(1), pp. 61-71

[2] Chin, W.W. (1998). Issues and Opinion on Structural Equation Modeling. *MIS Quarterly*, 2(2), pp. vii-xvi

[3] Christen, M. (2005). Cost Uncertainty is Bliss: The Effect of Competition on the Acquisition of Cost Information for Pricing New Products. *Management Science*, 51(4), pp. 668-676

[4] Davenport, T.H. (1997). *Information Ecology. Mastering the Information and Knowledge Environment*. Oxford University Press, N.Y.

[5] Dell'Aricca, G.; R. Marquez (2008). Can Cost Increases Increase Competition? Asymetric Information and Equilibrium Prices. *The Rand Journal of Economics*, 39(1), pp. 144-162

[6] Dewan S.; H. Mendelson (1998). Information Technology and Time-Based Competition in Financial Markets. *Management Science*, 44(5), pp. 595-609

[7] Hauswald, R.; R. Marquez (2006). Competition and Strategic Information Acquisition in Credit Markets. *The Review of Financial Studies*, 19(3), pp. 967-1000

[8] Jarvenpaa, S.; B. Ives (1993). Organizing for Global Competition: The Fit of Information Technology. *Decision Science*, 24(3), pp. 547-580

[9] Krishnan, R. (2005). The Effect of Changes in Regulation and Competition on Firms' Demand for Accounting Information. *The Accounting Review*, 80(1), pp. 269-287

[10] Lavon, G.; M. Todd (2011). Information Technology and Its Role in Creating Sustainable Competitive Advantage, *Journal of International Management Studies*, 6(1), pp.1-7

[11] Leckson-Leckey, G.; K.A. Osei; S.K. Harvey (2011). Investments in Information Technology (IT) and Bank Business Performance in Ghana. *International Journal of Economics and Finance*, 3(2), pp. 133-142

[12] Leidner, D.E.; J.J. Elam (1995). The Impact of Executive Information Systems on Organizational Design, Intelligence, and Decision Making. *Organization Science*, 6(6), pp. 645-664

[13] Lillrank, P. (2003). The Quality of Information. *International Journal of Quality & Reliability Management*, 20(6), pp. 691-703

[14] Loukis, E.; I. Sapounas (2008). The Effect of Generalized Competition and Strategy on the Business Value of Information Communication Technologies. *Journal of Enterprise Information Management*, 21(1), pp. 24-38

[15] Marchand, D.; W. Kettinger; J. Rollins (2002). *Information Orientation: The Best Link to Business Performance*. Oxford University Press, N.Y.

[16] Mendelson, H. (2000). Organizational Architecture and Success in the Information Technology Industry. *Management Science*, 46(4), pp. 513-529

[17] Melville, N; K. Kraemer; V. Gurbaxani (2004). Review: Information Technology and Organizational Performance. An Integrative Model of IT Business Value. *MIS Quarterly*, 28(2), pp. 283-322

[18] Nunnally, J.C. (1978). *Psychometric Theory*. McGraw Hill Editorial, N.Y.

[19] Oppenheim, Ch.; J. Stenson; R.M.S. Wilson (2004). Studies on Information as an Asset III: Views of Information Professionals. *Journal of Information Science*, 30(2), pp. 181-190

The improvement of customer relationship management process by investing on information technology (IT) in an organization

Mohammad Reza Moini[1], Milad Torfi[2], Marjan Abdeyazdan[3], Hamid Reais Ghanavati[4]

[1]Department of Computer Engineering, Mahshahr branch, Islamic Azad Universiy, mahshahr, Iran.
e-mail: rezamoini_it@yahoo.com
[2]Department of Computer Engineering, Mahshahr branch, Islamic Azad Universiy, mahshahr, Iran.
e-mail: miladtorfi@gmail.com
[3]Department of Computer Engineering, Mahshahr branch, Islamic Azad Universiy, mahshahr, Iran.
e-mail: abdeyazdan87@yahoo.com
[4]Department of Company Ghazalrh Mahshahr, mahshahr, Iran.
e-mail: hamid_raeis@yahoo.com

Abstract. The increasing development of IT produces modern ways of relationship among organizations, institutions, and commercial agencies with customers. The customer relationship management is a process that helps organization to achieve a better prospective of its person` commercial and ethical behavior. This significant process is established by IT. By closely considering different aspects, this article investigates the effect of IT on the process of the customer relationship management. The essential substructures for having a good reciprocal connection between organization and customer also investigated. Therefore, by recommending a structure, an IT instruments, applied as an input, are used to develop the process of CRM. And, for example, the decreasing in costs will be functioned as one of effective output. Using this method, the organizations can benefit from applying IT in their marketing and different services and they easily can fulfill their needs to information and connections.

Key words: IT (information technology) - customer – organization - the customer relationship management (CRM)

1. Introduction

Today researchers and the marketing staffs pay a lot of attention to IT.A group believes that the term IT is firstly invented by Wiles and Luit in 1958 when it used to show the role of computer in supporting decisions and processing information in an organization. Other one contents that this term is coined in 1970 to referring to work out information by computer technology.
A lot of managers believe that the selection of products should be based on important factors such as IT. Because of these potentials IT is the most powerful factor in economical and social changes, and it is predicted that it will continue its progress in all different aspects. On the other hand, organizations have conceived that customers are their most valuable properties and look at the relationships with their customers as a profit interchanges and reciprocal connections.

In toady's business, the managers have comprehended that the costumers are in the center of marketing and the organization success depends on having an effective relationship with them. Regarding the global changes, the organizations should try to precede their customers' needs, and they should also give their customers continued and invaluable services. Thus we can say that IT will influence institutions and commercial agencies. By collecting data about customers by IT, organizations and commercial agencies, will establish an important background for their marketing, selling and servicing activities. They process these collected data by IT instruments and convert them to suitable information. And because of this reason a lot of organizations use IT to resist in the today's competitive marketing. Working on the effectiveness of IT in marketing, Borke and his colleagues emphasize the importance rearranging the use of web technology in the new marketing. The emergent of internet and its commercial application in1994 globally put a heavy responsibility on IT to convey information to customers and fulfilling their needs. Nowadays IT helps marketing by systems which can aid companies to follow the customers' interchanges and allow the staffs to quickly retrieve the customers' information. This process is called customer relationship management (CRM). Accordingly the customers' needs and expectations have to be identified and slicked in the staffs minds, and through continuous communication, they should attempt to increase their expectations.

Customer relationship management is not a traditional commercial philosophy but it was again born by development of IT. Restructuring the trade system in a complicated digital environment needs a careful programming, planning and executing a customer –oriented method. Accepting a customer –oriented strategy necessitate CRM.

In this article, besides presenting IT and CRM features, the IT function on CRM according various writers is also investigated. Then a recommended structure in which the components of IT make its input is presented. Also the advantages and potentials of the output are researched.

2. IT and the significant of investing on it

Investing in the IT domain is one of the important subjects in all organizations and its usage in the small productive industries dated to many years ago. Recently it has made different innovations in trade and marketing. For being successful in a compiutitve environment, companies should simultaneously raise the qualities of their goods and lower the prices. So they need IT. This technology can be either used in producing or management processes.
In many cases investing in this domain can lower the expenditures and costs. IT is basic element to empower the organization and an enduring competitive source in

commercial affairs. IT, also, increase the benefits of an organization. Investing on IT has an either tangible or intangible effects on an organization (it illustrated in figure 1).For reaching this end, an organization requires to have a combination of innovative IT, an effective trade processes, a better data management, and a new potential innovative labor strategy.

As a modern communicative means, IT empowers the organization to achieve its ends and remarkable grows by fundamentally planning the business and commercial processes and creating new innovative ways to connect the company to customers, shareholders and potential providers. The other benefit of IT are decreasing the period of expanding the production of new products, benefiting from the staffs massive cooperation, and extending the staffs creative domain through close relationship with customers , distributers and partners.

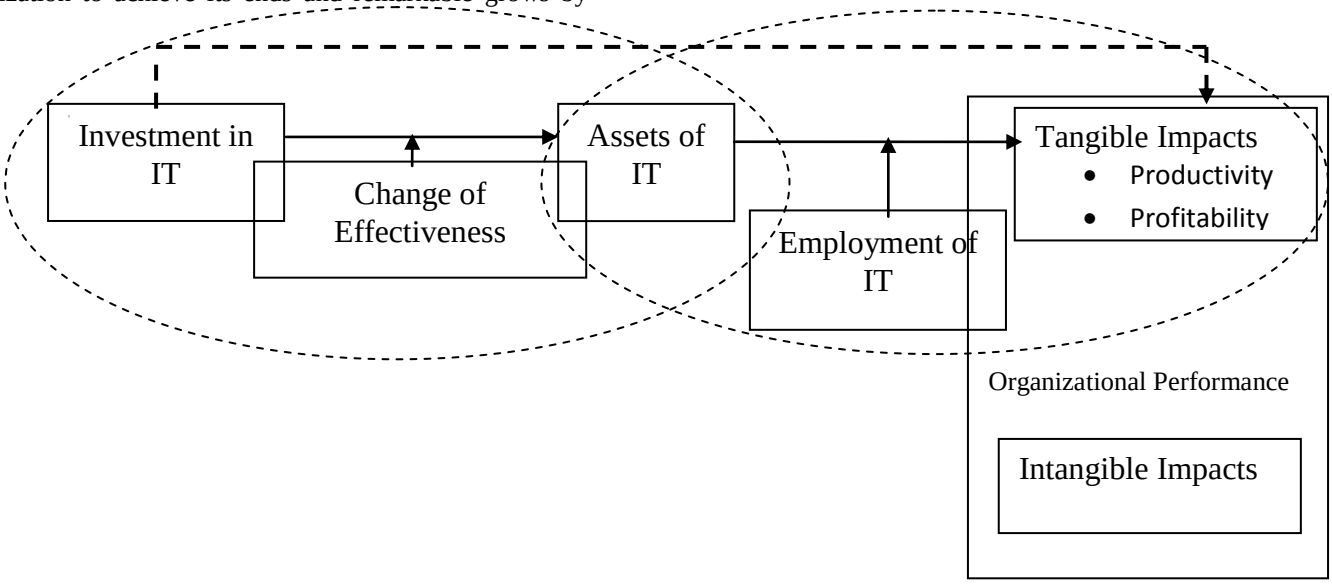


Figure 1: A perceptional model of IT investment impacts on organization productivity [10, 11]

3. Customer relationship management

Today in current business satisfying the customer's needs have a vital role within an organization ends. The senior manager know that their successfulness in achieving high level ends depend on customers satisfying. Customer management is not a new thing, rather it is distinguished from other domain in the middle of 90s account for expanding in customers management techniques which are initiated by IT for tracking the customers activities. First Provatiar and Shes (2001) made the basics of CRM synonym with the concept of communications inn the marketing section.

They equaled the CRM expanding for variations in trading within IT especially in innovation in common similarities between companies and their relations with customers. O they unified the quality philosophy with attempts of decreasing the costs. CRM is embedded in customer data and it is facilitated by IT. Actually CRM is modern means for analyzing the customer's data that is provided by connection points, so it presents a general outlook of any customer. In some economical agencies CRM is just technology which leads in improved activities of an organization by expanding information banks, automation means and also connecting the selling duty with marketing. An organization should be able to present its activities numerically so that it could perceive its processors since not to perceiving the processes will not lead in improvement. Assessing the processes based on economical scales is the only way to improve the relationship with customer.

CRM is just a simple instance of marketing connections which aim at preserving the customers and it also make a durable communications that improve the customer values for companies. In the customer –oriented process the customer is the manager and the producers produce what the customer needs. These kinds of organization by moving from product –oriented to customer –oriented way emphasize the customers' needs. They strengthen the connections through using IT and in the suitable time they produce high quality goods.

CRM is a planned process for collecting data about customers and using them in business activities. The organizations that assess their customers ` attitudes and convert them to key processes will produce products according customer needs. So CRM as a program for connecting customers relies on uses IT. Such programs besides increasing loyalty, they have a lot of applications.

4. The abilities of IT in improving CRM process

IT systems are used for learning, storing accessing the customer information and it is also used to analyze itself. Technically is an ability that to have a close relationship with customer, to analyze the customer data for having a prospective from a customer. It s supposed that the business affaires, by applying IT, achieve thieve their ends. Chirco and kafman(2000) believe that an organization can benefit from IT to reach its competitive advantages from the organizational available resources.

IT instruments for reaching better harmony have been selected based on business activities. Therefore, CRM always require a skillful IT specialist. IT helps the organization to anticipate its success and the customer has bigger part than market.
Also collecting, transferring, and unifying the customers data needs IT. Actually IT is the vital but not enough condition in CRM since it needs other instrument. Teo and his collogues (2006) present a general structure for CRM

which is based on three operational, analytical and cooperative elements in an 43organization. Also the other three elements, i.e technology, business, and customer make CRM communicative ring (figure 2).

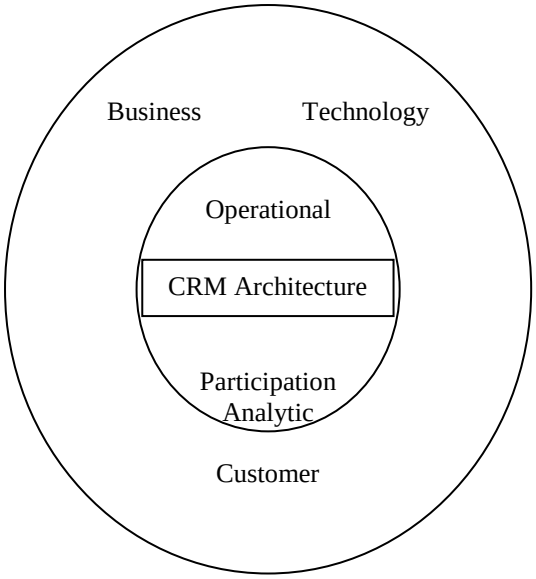


Figure 2: A holistic framework for CRM [22]

IT support for CRM change according complication, administration difficulty, and the range of customer supporting. Chung and Van Co (2009) assert that IT is a combination of marketing attempts and business and trade processes which allow the companies differently understand their customers. According to their pint of view, CRM uses IT and knowledge to unify marketing, selling and services; and it also make some change to improve the customers' satisfaction, increasing the number of loyal customer and finally raising the earning of customers. Brinjofston and Heat recommend that the credit and value of IT should be assessed by invisible scales like improving the quality, giving serves to customer, and expand new product .It can be said that IT and CRM have a reciprocal relationship.

CRM administration is a sign of using a united strategy of customer data and a constant plan by applying IT. Changing the programs is a result of a kind of new business and trade processes like CRM which is uniquely is guided by IT. Particularly, using IT leads in a lot of customers by applying CRM. All customers like unifying all the parts after establishing CRM.

5. Presenting the recommended structure

Although CRTM is not a new thing but considering the development in IT it has become more practical than before. This technique tries to collect and analyze information about customers and by using effective IT make relationship with customers and establish a customer –oriented context that will lead in many more long –term benefits. In the former section the influence of IT on CRM is investigated according different writers but no outcome from these significant and key potentials which include the other subtle and effective elements, have been recommended. So in this section, we, using all the potentials of IT which increased daily, intend to apply a structure of IT as an effective way for reaching the organization ends and to mechanize the processes.

In this recommended structure the different IT instruments, which are used by companies and commercial agencies, are defied as an input. The other communication means which are used in collecting data include E-mail, short messages, telephone and fax. In the CRM unit the data obtained from different sources, are analyzed and processed by the experts. This unit operates as a ship watchman. This watchman by knowing the needs and expectations of the customer and by transferring them to the producer improves the goods quality, and leads in change in the organization processes. In fact, the customers themselves dictate the organization policies and strategies. It means that the organizations should harmonize themselves with customers.

It should not be forgotten that the organization ability in selling and its willing to use IT for reaching the intended end of an innovative CRM is an essential circumstances.

In fact, if it is used properly, IT can be helpful to effectively manage customers and establishing a strong and loyal relation with them. The recording and analyzing the data of the customers help agencies to identify their customer easily and undoubtedly IT is essential for administrating CRM.

There is a valuable result in the out put stage. The outputs that companies need to apply all the available instruments to improve them include: lowering the costs, quicken the communications, advantages of constant competition, remote access, increasing the personnel proficiency.

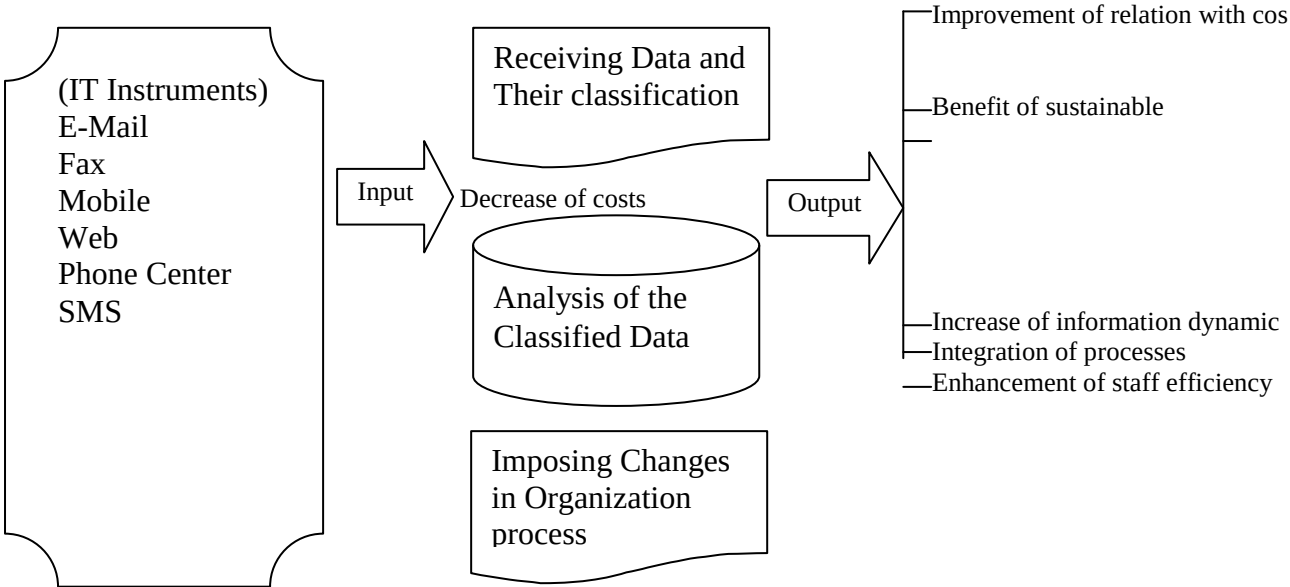


Figure 3: A proposed framework for CRM based on employment of IT

6. Conclusion

Nowadays, IT helps companies to obtain customers who are far from them, thus, if they manage to select their product, IT can help them to buy it and also it helps them to survey and criticize about the products**.** IT can lead in direct relationship between companies and customers and help to interchange the information; and generally IT is a good chance for companies.

The significance of IT for companies has remarkably increased and it needs to devolve more and more .However, IT is not used properly in small and medium companies and is lower than the big ones. Actually the organization, using IT in their interchanges, can have 360 degrees vision to their customers and in future can have better relationship with them.

What is understood from the effect of IT on companies is the importance of paying much more intention to this part of business. As an example developing the backgrounds in relation to skills and standards should be emphasized and provide all the actual and legal conditions. AN origination should train their staffs to use IT if it wants to follow the quick development of IT.

Nowadays simultaneous with IT development the inside movements of the organization and decreasing the costs and electronic trade encourage the organizations to use techniques to attract and preserve customers. Before making any decision about applying any IT strategy, the organization should assess this process, not doing so the some long and short –term strategic benefits will be remained unidentified.

IT development through different ways helps to improve the relationship between companies and customers and comprehending the speed and development in electronic trade is very important. As an example the companies can connect to their customers by electronically presenting the goods I way that the characteristics of the goods are described. In the other word, in IT business context the customers need to have all information about sellers, products, services and goods.

The companies through IT, storing and analyzing information about customers can produce goods according the customers needs and expectations. CRM as one of information systems in organization can easily cover all the information relation needs through combination of IT, marketing and services.

The other effects of using IT and applying CRM are:

-collecting and analyzing data about customers
-determining the customers behavior
-reciprocal and constant relationship with customers
-giving special services t special customers like transferring the product to them
-generating and developing the service level patterns

Even though CRM administration needs investing on IT, it will have beneficial output as a result. Supporting development of marketing to customers depends on using effective IT. The companies that use IT structures to preserve their customers should increase this orientation.

In sum CRM strategy of using IT, marketing, selling services to customers, human resources can increase the financial benefits and the relationship between customers and organization.

At the end of this article, because of a lot of unsuccessful applying cases of CRM, it should warn not to make any unstudied decision on CRM investing before considering the all aspects of using IT.

References

[1] Grauer, M. 2002. Information technology. In International Encyclopedia of Business and Management, edited by M. Warner, 3014-3024. Australia: Thomson learning.
[2] Winter, Susan S.; and S. Lynne Taylor. The role of information technology in the transformation of work. In Information Technology and Organizational Transformation: History, Rhetoric, and Practice, edited by Joanne Yates and John Van Maanen, 2001,pp. 7-33.
[3] Bahrami, M.2010-2011. *A study in the effect of IT on relationship management and present a recommended pattern.* First IT conference in Sanadaj, Iran.2010-2011.
[4] Nguyen, T., Sherif, J., Newby, M., Strategies for successful CRM implementation,Information Management & Computer Security Vol. 15, pp. 102-11, 2007.
[5] Burke,R.R.,Rangaswamy,A. and Gupta,S, Rethinking marketing research in the digital word, working paper, eBusiness Research Center, Pennsylvania State University, University Park,PA, 1999.
[6] Cortada, J, W., How Computers Changed the Work of American Manufacturing, Transportation, and Retail Industries. Oxford University Press, Oxford, New York, 2004.
[7] Wilcox, P., Gurau, C., Business modelling withUML: the implementation of CRM systems for online retailing, Journal of Retailing and Consumer Services,vol. 10 ,pp. 181 191, 2003.
[8] Mephkee, C., and Ruengsrichaiya, K., Information and communication technology (ICT) for development of small and medium-size exporters in East Asia: Thailand. Santiago, Chile:
UnitedNation,
http://www.eclac.cl/comercio/IT_SME/documentos /SW-66 hailand.pdf (accessed 15 Jun. 2006), 2005.
[9] Rahimifard, S., R. W. Bagshaw, S. T. Newman, and R. Bell., IT tools to improve the performance of metalworking SMEs, International Journal of Production Research vol. 40,pp. 3589-3604, 2002.
[10] P. Weill, The relationship between investment in information technology and firm performance: a study of the valve manufacturing sector, Information Systems Research 3 (4) (1992) 307 333.
[11] C. Soh, M.L. Markus, How IT creates business value: a process theory synthesis, Proceedings of the 16th International Conference on Information Systems, Amsterdam, The Netherlands (1995) 29 41.
[12] Minami, C., Dawson, J.,The CRM process In retail and service sector firms in Japan:Loyalty development

and financial return, Journal of Retailing and Consumer Services,vol. 15, pp. 375 385, 2008.
[13] Parvatiyar, A., Sheth, J, N., Customer relationship management: emerging practice, process, and discipline, Journal of Economic and Social Research,vol. 3,pp. 1 34, 2001.
[14] Wehmeyer, K., Aligning IT and marketing the impact of database marketing and CRM , Journal of Database Marketing & Customer Strategy Management,vol. 12,pp. 243 256, 2005.
[15] Swift, R, S., Accelerating Customer Relationships: Using CRM and Relationship Technologies, Prentice-Hall, Upper Saddle River, NJ, 2001.
[16] Batislam, E, P., Denizel, M & Filiztekin, A., Empirical validation and comparison of models for customer base analysis, International Journal of Research in Marketing, 24(3),pp. 201 209, 2007.
[17] Rust, R, T & Verhoef, P, C., Optimizing the marketing interventions mix in intermediate-term CRM, Marketing Science,24(3),pp. 477 489, 2005.
[18] Goodhue, D, L., Wixom, B, H., Watson, H, J., Realizing business benefits through CRM: hitting the right target in the right way, MIS Quarterly Executive,vol. 1,pp. 79 94, 2002.
[19] Chircu, A, M., Kauffman, R, J., Limits to value in electronic commerce-related IT investment,JManage Inf Syst;17(2)pp. 59 80, 2000.
[20] Chang, H, H., Wen Ku, P., Implementation of relationship quality for CRM performance:Acquisition of BPR and organisational learning, Total Quality Management & Business Excellence, Vol. 20,pp. 327 348, , 2009.
[21] Minami, C., Dawson, J.,The CRM process In retail and service sector firms in Japan:Loyalty development and financial return, Journal of Retailing and Consumer Services,vol. 15, pp. 375 385, 2008.
[22] Teo, S, H., Devadoss, P, L., Pan, S., Towards a holistic perspective of customer relationship management(CRM)implementat on: A case study of the Housing and Development Board, Singapore, Decision Support Systems,vol. 42,pp. 1613 1627, 2006.
[23] Brynjolfsson, E., Hitt, L, Beyond the productivity paradox, Commun ACM;41(8)pp. 49 55, 1998.
[24] Yang, S., Rhee, J., Study of the wireless/wire integration CRM Gateway for the effective application of Event CRM for small and SME, Computers & Industrial Engineering, vol. 57,pp. 571 579, 2009.
[25] Van Bruggen, G, H & Wierenga, B., When are CRM systems successful? The perspective of the user and the organization, Erasmus Research Institute of Management Report Series Research in Management, pp. 1 50, September 2005.

SESSION

APPLICATIONS AND RELATED ISSUES

Chair(s)

TBA

Adaptive Traffic Light Controlling Methodology Using Connected Vehicles Concepts

M. Arafat[1,2], U. Mohammad[1], N. Al-Holou, Ph.D.[1], M. A. Tamer[1], and M. Abdul-Hak[1]
[1]Department of Electrical and Computer Engineering
University of Detroit Mercy, Detroit, Michigan, USA
[2]Department of Logic of Usage, Social Sciences and Information (LUSSI)
TELECOM Bretagne, Brest, France

Abstract— *Despite the fact that they play a major role in organizing traffic flow, traffic lights can sometimes introduce significant delays and cause severe traffic congestion, especially during peak demand hours. However, based on the connected vehicles concepts proposed by USDOT, where vehicles will be able communicate with infrastructure and with other vehicles, adaptive traffic light controlling methodologies should provide the best performance for such ITS applications, and significantly help in reducing waiting times and congestion. Although it has been well studied by many in the research community, adaptive traffic light controlling algorithms have not been widely studied in a V2X-based environment.*

In this paper, a methodology for adaptively controlling traffic lights, using V2X exchanged messages, is proposed and evaluated. Compared to a relatively smart pre-timed algorithm, which defines different cycle lengths and phase splitting for different times of the day, the proposed methodology outperforms this pre-timed one for many measures of effectiveness such as control delay introduced by the traffic light, as well as queue length, waiting and traveling times.

1. Introduction

The first four-way, three-state traffic light was invented by police officer "William Potts" in Detroit, Michigan in 1920. Ever since then, traffic lights have been one of the most important means for traffic management. In fact, they have proven during the past century their efficiency in organizing traffic flow, and reducing traffic jams. The concepts of connected intersection control, and automatic traffic lights are as old as the idea itself. The first interconnected traffic signal system was installed in Salt Lake City in 1917, with six connected intersections controlled simultaneously from a manual switch. Automatic control of interconnected traffic lights was introduced March 1922 in Houston, Texas [16].

When the need for intelligent transportation systems arose, traffic lights received a considerable amount of attention, and many approaches to achieve adaptive traffic lights were proposed. These included, but were not limited to, metal detectors and computer vision. However those systems lacked accuracy because they were based on the accuracy of sensors, or the robustness of the computer vision algorithm used. After the connected vehicles concept was proposed, in which every vehicle shares information about its status with other vehicles and infrastructure, a wide range of new possibilities arose, and one of them was adaptive traffic light systems based on accurate real time traffic information collected from the vehicular network.

The problem of real-time adaptive traffic lights has been well covered in the literature with two main streams of interest, one focused on the algorithm used to determine cycle length and green phases timing, regardless of the means used to collect traffic data; while the other focused on introducing new ways of collecting traffic data, with the help of a simple algorithm to present performance evaluation and results. B. Zhou *et al.* propose an adaptive traffic light control algorithm that uses traffic data collected from a wireless sensor network to determine sequence and length of the traffic light phases [17]. The authors assume the intersection can only be in one of 16 cases, and they use a mathematical model to determine the intersection's next case and the period over which it should persist. On the other hand, D. T Dissanayake *et al.* propose an algorithm for vehicle detection based on a Magneto-Resistive sensor [18]. Also, K. Al-Khateeb *et al.* propose a real-time dynamic traffic light sequence determination algorithm, but this time using RFID technology to collect real-time traffic information [19]. Although some of the published research present a relatively easy algorithm to determine traffic light cycle length and phases timing, others present very complicated algorithms that incorporate the learning abilities of artificial neural networks, with the decision making of fuzzy expert systems such as the work presented in [20].

Despite the high level of attention to adaptive traffic light systems, only a few papers considered the connected vehicles concepts (i.e. vehicle-to-vehicle and vehicle-to-infrastructure (V2V/V2I) communications) as the source for the real-time traffic information. M. Ferreira *et al.* present a new concept of traffic management at intersections, using only V2V communications [21]. The proposed algorithm does not require either roadside equipment (RSE), or a traffic light. The algorithm is only based on communication

between vehicles at the same intersection, and the traffic light is replaced with an internal traffic light presented on a display in each vehicle. Another V2V/V2I utilization was presented by V. Gradinescu *et al.*, in which V2V/V2I messages are used to collect real-time information about the traffic conditions around the intersection, and then a simple algorithm based on the well known Webster's equation is used to determine cycle length as well as green time splitting [22].

In this paper, an adaptive traffic light controlling methodology using V2V/V2I communications is proposed and evaluated. Traffic information will be collected from V2V/V2I messages exchanged with cycle length and green times determined using an algorithm based on Webster's equation. The remainder of the chapter is organized as follows, section 2 summarizes the different classifications of traffic light control systems known in the literature; section 3 presents the algorithm used to determine the traffic light's cycle length and green times; section 4 describes the simulation model used in detail; evaluation and validation results are discussed in section 5; and a conclusion along with future possible work puts this chapter to an end in section 6.

2. Classification of Traffic Light Controlling Systems

A traffic light can be defined by three major elements which are, cycle length, green time splits, and relation to the surrounding environment. Accordingly, traffic lights can be classified into three main categories: pretimed, actuated and adaptive.

In pretimed traffic lights, cycle length and green time splits are pre-determined before the traffic light is put to operation. In addition, the traffic light does not respond to any sudden changes in the surrounding environment. This is the most basic and simple form of a traffic light. Further enhancements were done by defining different programs (cycle length and green times) for different times of the day, or day of the week. Historical data about traffic flow were used to find peak hours and assign suitable cycle length and green times accordingly.

Actuated traffic lights form an enhanced version of pretimed traffic lights. In actuated traffic lights, cycle length and green times are pre-timed, however the traffic light's ability to respond to surrounding environment events is introduced by adding sensors on some, or all, the controlled roads by that traffic light. Thus the main difference between pretimed and actuated is the ability to respond to some events from the surrounding environment. For example, in an intersection comprised of a major road crossed by a secondary road, an actuated traffic light can be used to extend green on the main road as long as no traffic is present at the secondary road; in case of incoming traffic on both directions, the traffic light will work based on the pretimed program previously defined for that time of the day and the day of the week.

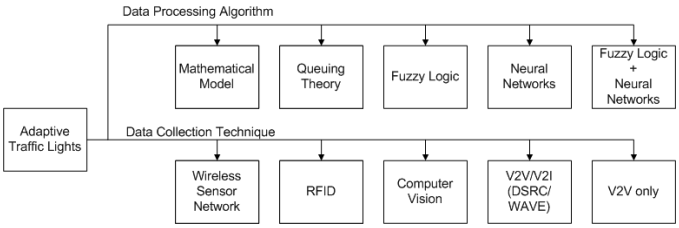


Fig. 1: Different Categories of Adaptive Traffic Lights

The third category of traffic lights is the adaptive one, in which cycle length and green times are calculated based on traffic data collected in a real-time manner from the surrounding environment. Several sub-categories can be defined based on traffic data collection technique, and the algorithm used to calculate different parameters and assign green and red lights. Figure 1 shows the different sub-categories of adaptive traffic lights based on data collection technique and data processing algorithm.

3. Algorithm

Algorithms for controlling adaptive traffic lights can be simple (simple mathematical model), or very complicated (combination of fuzzy logic and artificial neural networks). Webster's equation is a tool used to determine the optimal cycle length for a traffic light according to traffic flow information and lost times such as yellow times and all-red times. Traffic flow information is usually taken from historical data and fed into the equation. However, since vehicles can communicate with each other and with the infrastructure using V2V/V2I communications, the data can be collected from the incoming vehicles themselves, and can be fed into the equation to come up with an optimized cycle length and green times for the current situation which will result in optimized performance and increased throughput of the controlled intersection. Equation 1 shows Webster's equation in which demand levels are represented by a factor called *critical flow ratio* which is the ratio of the current flow rate captured from the exchanged messages, over the road's saturation flow rate.

$$C_0 = \frac{1.5 \cdot L + 5}{1 - \sum_{i=1}^{i=n} y^i} \quad (1)$$

where: C_0 is optimal cycle length [sec], L is lost times (yellow + all-red) [sec], y^i is critical ratio of road segment group i, and n is number of road segment groups; where the road segment group is the group of road segment on which

incoming flow can access the intersection simultaneously.

The idea is to calculate cycle length and green times splits every new cycle so the system can dynamically adapt to changes in the input traffic flow. This algorithm uses information from messages exchanged between vehicles, those messages contain information about the vehicle's position, speed, acceleration, and many different parameters. Since these messages are received in every simulation step, a special procedure for accumulating the information over a complete cycle length is needed. For that, a procedure from one of the published papers was imported and used here. The procedure is proposed by S.-F. Cheng *et al.*, and can be found in the appendix of their published paper about an algorithm for coordinated traffic lights [23].
The procedure calculates—for each road segment group—the estimated flow as follows:

$$v^i = f^i + 4q^i \tag{2}$$

where v^i is the estimated flow for road segment group i, f^i is the exponentially smoothed average incoming flow on road segment group i, and q^i is the exponentially smoothed average size of the standing queue on road segment group i.
Both average incoming flow (f^i) and average size of the standing queue (q^i) of road segment group i are obtained by periodically performing the following exponential smoothing updates:

$$f^i := 0.75f^i + 0.25f^i_{in} \tag{3}$$

$$q^i := 0.9q^i + 0.1\hat{q}^i \tag{4}$$

where f^i_{in} is the number of vehicles flowing into road segment group i during the interval between smoothing updates (one simulation step), and $\hat{q}^i$ is the size of standing queue on road segment group i during the same interval.

Then the critical flow ratios are calculated. Those values are used as a measure to represent the relative congestion of each road segment group, and thus help in the calculation of cycle length and green times. Critical flow ratios are calculated—for road segment group i—as the ratio between the estimated flow and the saturation flow rate:

$$y^i = \frac{v^i}{m \times s^i} \tag{5}$$

where v^i is the estimated flow calculated in equation (2), s^i is the saturation flow rate of one road segment, and m is the number of road segments inside of a road segment group i.

In order to use Webster's equation to calculate cycle length, a final parameter must be defined, which is lost times. For this work, lost times are only limited to yellow times. There are two yellow phases for the considered intersection. To determine the time in each of those yellow phases, a well known rule in traffic design was used. This rule determines yellow time for an approach of an intersection (in seconds) according to the speed limit(in miles per hour (mph)) on that approach. According to this rule, 1 second of yellow time should be scheduled for every 10 mph in the maximum speed allowed. In our design (explained later in section 4), the speed limit on all approaches is 40 mph, and thus one phase of yellow time is 4 sec. In conclusion, L in Webster's equation will be replaced by 8 seconds.

Now Webster's equation can be used to calculate the optimal cycle length C_0 according to the estimated flows. Also green times can be calculated according to critical ratios calculated in equation (5) as follows:

$$g^i = \frac{y^i}{Y}(C_0 - L) \tag{6}$$

where g^i is the green time that should be associated with the road segment group i.

Some limitations should be introduced to the algorithm to ensure the optimal result. For example, minimum green time, which is the minimum amount of time required by a pedestrian to cross the road segment with an average speed 4ft/s should be respected. Also, a minimum and a maximum cycle length should be respected, where the minimum cycle length is the sum of two minimal green times with two yellow times, and the maximum cycle length is normally 1.5 C_0 for a C_0 calculated in moderate conditions.

The algorithm has also been extended to deal with special traffic cases. There are two cases that are covered by the programmed algorithm and those are: *a*) eliminating tedious waiting times on red by switching the traffic light automatically to green when the opposite direction has no demand; and *b*) extending a finished green phase when the opposite direction has no demand.

4. Simulation Model Design

The simulation scenario that will be used to validate the concept needs to be as realistic as possible. To build such a scenario, the following parts of the scenario should be addressed and defined carefully: *a*) the road network; *b*) traffic flows that will run through the network; and *c*) the routing algorithm used to route vehicles between their respective origin and destination. Since this simulation scenario will be targeting the evaluation of the performance of adaptive traffic light controlling algorithms, only a single intersection is needed, and thus routing will not have an essential role in the evaluation process.

The road network consists of an isolated 4-way intersection, with each approach having two directions, as

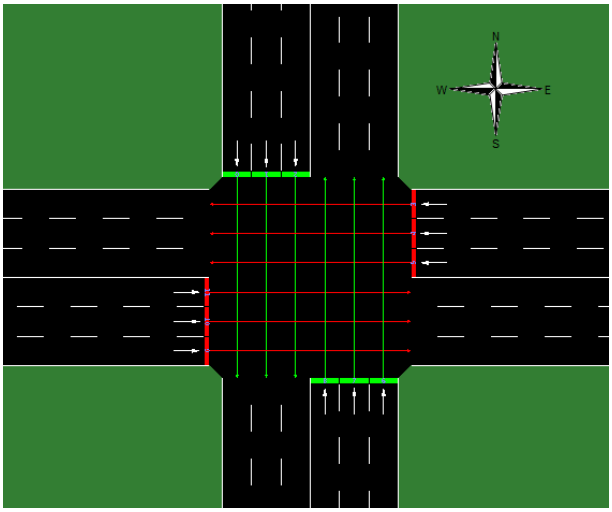


Fig. 2: Proposed Intersection

Table 1: Vehicles Types in the simulation

Vehicle Type	Max Speed [m/s]	Acceleration [m/s^2]	Length [m]	Probability [%]
Typical	70	2.68	7.5	49
Fast	80	3.83	7	19
Slow	60	1.92	6.5	22
Van	60	2.44	10	10

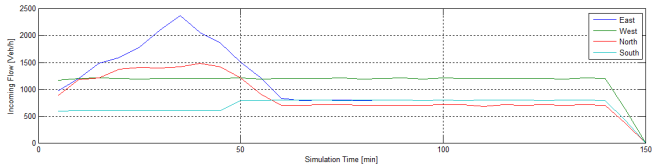


Fig. 3: Incoming Flow over Simualtion Time for all scenarios

illustrated in Figure 2. Since similar work had been done in the literature by V. Gradinescu *et al.* [22], we tried to reproduce their model and thus acquire the opportunity to validate our simulation by comparing both results. However, in their paper, the model used was not fully specified. They provided the shape of the intersection, number of lanes and input flow rates on each approach, but the model lacked detailed description of the intersection such as left/right turns, speed limit on different approaches, vehicle-based parameters (max speed, acceleration, deceleration, length), the distribution of different types of vehicles, and the pre-timed cycle length and green times of the traffic light controlling the intersection. In other words, many elements that define the simulation scenario were unspecified, and thus it is not possible to compare both results.

The road network used for this scenario is an isolated 4-way intersection—and by *isolated* we mean that the effect of adjacent intersections was not considered—with each approach having two directions, and with each direction having three lanes. No left or right turns are allowed at the intersection. The speed limit on all lanes is 40 mph $\approx$ 17.89 m/s. Vehicles participating in the scenario were assumed to be of four types: *Typical*, *Fast*, *Slow* and *Van*, Table 1 specifies the different parameters and distribution for each vehicle type.

Input flow of vehicles into the intersection should be defined very carefully to represent situations such as peak hour demand. For this evaluation process, the used input flow definition was inspired by the one used by V. Gradinescu *et al.* in [22]. It assumes the simulation covers almost three hours of real life, during which a peak demand will occur. This helps in understanding the performance of the evaluated algorithm in different demand levels. The input flow on different approaches and their variations over the simulation time are illustrated in Figure 3.

The pre-timed traffic light cycle definition used is the one generated by default by SUMO. It gives 31 simulation seconds to each approach as green time and 4 simulation seconds as yellow time.

5. Evaluation and Validation

The simulation scenario simulates three hours of real life during which a peak demand will occur, and then the demand will go down to a normal level. The measure of effectiveness (MOE) used to evaluate the performance of the algorithm is basically the *average control delay* which is a widely used MOE for the evaluation of traffic light controlling algorithms, and it is defined as the difference in travel time for vehicles when travelling down a road with the traffic light controlling that road and when there is no traffic light controlling that road.

As illustrated in Figure 4, the simulation shows that the used algorithm outperforms the pre-timed traffic light over the entire simulation time. In addition, the algorithm improves the recovery time of the intersection after a the peak demand.

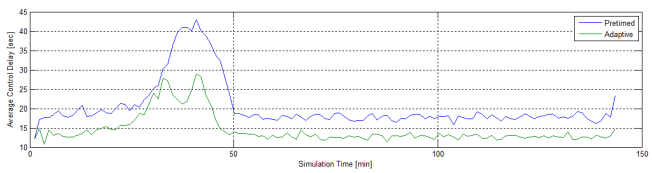


Fig. 4: Average Control Delay for the algorithm compared to that of conventional pre-timed control

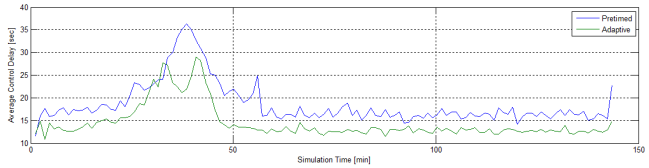


Fig. 5: Average Control Delay for the algorithm compared with a "smart" pre-timed control

Moreover, after evaluating different cycle lengths for the pre-timed program, it was noticed that long cycle lengths are suitable for high demand situations, and short cycle lengths are suitable for low demand periods, thus a smart pre-timed program was evaluated. The traffic light starts with a short cycle length (60 min) up to minute 25 were it switches to a longer cycle length (80 min) with a fixed yellow times. Then, when the simulation reaches minute 55, the traffic light switches back to the old timing. The results for this smart pre-timed program are shown in Figure 5. The adaptive algorithm used outperforms the "smart" pre-timed program and provides lower control delays over the entire simulation time.

Other MOEs have been evaluated such as queue length, waiting time, travel time, and reductions in emissions and fuel consumption.

Figure 6 shows the results of evaluating queue length in front of the traffic light. Although the adaptive algorithm increased queue length on northbound lanes during a portion of the peak demand period, the algorithm succeeded in reducing queue length in front of the traffic light over the entire simulation time. This indicates that the adaptive algorithm has succeeded in reducing congestion at the intersection.

In terms of waiting and travel time, the adaptive algorithm improves the over all performance although small delays are introduced on northbound lanes while reducing them on eastbound lanes. Figures 7 and 8 show the simulation results of evaluating those parameters on every bound.

6. Conclusion

In this paper, a methodology for adaptively controlling traffic lights in a connected vehicles environment was proposed and evaluated.

The algorithm was based on the well-known Webster's equation, which determines the optimal cycle length and green times based on the demand level in each direction. The algorithm used messages exchanged between the vehicles and the traffic light (V2I) to estimate demand level and thus calculate a suitable cycle length and green times splits. Many parameters were considered in the algorithm such as minimum and maximum cycle length, and minimum green time for one direction. Also, the algorithm was able to react to special event at the intersection such as switching the traffic light to green for a direction waiting on red while there is no demand on the opposite direction, and extending the green phase beyond the calculated value for one direction as long as the opposite direction has no demand while respecting maximum acceptable waiting times for pedestrians.

iTETRIS, the open source simulation platform, was used to simulate the algorithm. The simulation scenario was three hours long. By the end of the first hour, the demand on the intersection reaches a peak level, and then gradually decreases back to a low level. The evaluation proved that the proposed algorithm outperformed simple and smart pre-timed controlling algorithms. Average control delay—which is a widely used MOE for evaluating traffic light performance—was mainly used to compare the proposed adaptive algorithm with the pre-timed examples, and the comparison showed that the adaptive algorithm succeeded in reducing average control delay, and even enhancing recovery time after the event of a peak demand. Further MOEs were evaluated such as queue length, waiting and travel time, and results show improvement in all of them.

Acknowledgements

This research has been funded by The Michigan Ohio University Transportation Center, the U.S. Department of Transportation and the University of Detroit Mercy- Internal Research Fund through UDMPU.

References

[1] U.S. Department of Transportation, Research and Innovative Technology Administration, Bureau of Transportation Statistics, Transportation Statistics Annual Report 2009 (Washington, DC: 2009) Available: http://www.bts.gov/publications/transportation_statistics_annual_report/2009/pdf/entire.pdf

[2] The Intelligent Transportation Society of America, "NORTH AMERICAN INTELLIGENT TRANSPORTATION SYSTEMS: ITS INDUSTRY SECTORS AND STATE PROGRAMS" (Dec, 2009) Available: http://www.itsa.org/itsa/files/pdf/Market%20Data%20Analysis%20Project%20-%20Phase%201%20Report%20(Final).pdf

[3] Rehunathan, Devan Bing, "Federating of MITSIMLab and ns-2 for realistic vehicular network simulation" in Proc. Mobility Conference 2007 - The 4th Int. Conf. Mobile Technology, Applications and Systems, Mobility 2007, Incorporating the 1st Int. Symp. Computer Human Interaction in Mobile Technology, IS-CHI 2007, p 62-67, 2007

[4] H. Park, Miloslavov, Lee, Veeraraghavan, B. Park and Smith , "Integrated Traffic/Communications Simulation Evaluation Environment for IntelliDrive^{SM} Applications Using SAE J2735 Dedicated Short Range Communications Message Sets" to be presented at the 2011 Annual Meeting of the Transportation, University of Virginia, United States, (Nov, 2010)

[5] B. Liu, B. Khorashadi, H. Du ; D. Ghosal,C. N. Chuah, M. Zhang, "VGSim: An integrated networking and microscopic vehicular mobility simulation platform" IEEE Communications Magazine, v 47, n 5, p 134-141, 2009

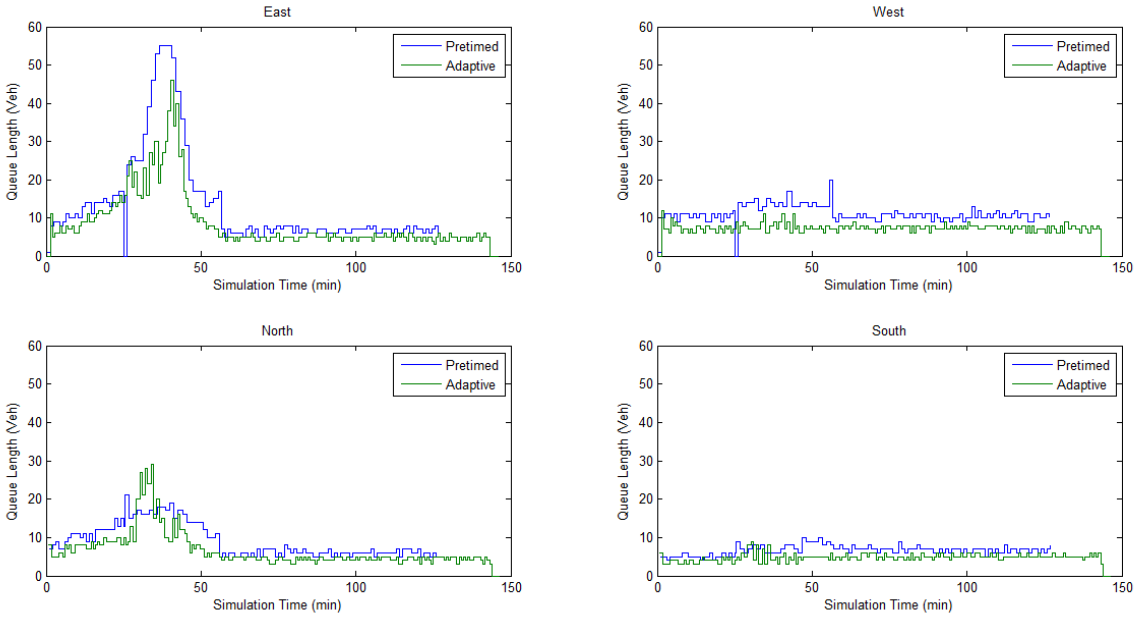


Fig. 6: Queue Length variations in front of the intersection on all bounds

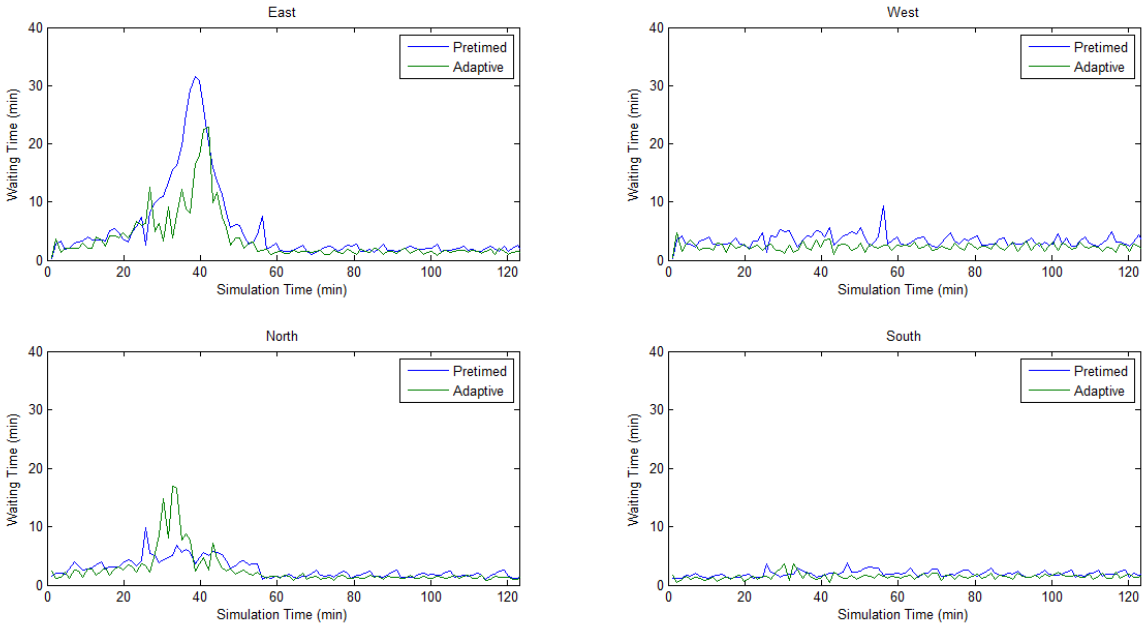


Fig. 7: Waiting Time variations over simulation time on all bounds

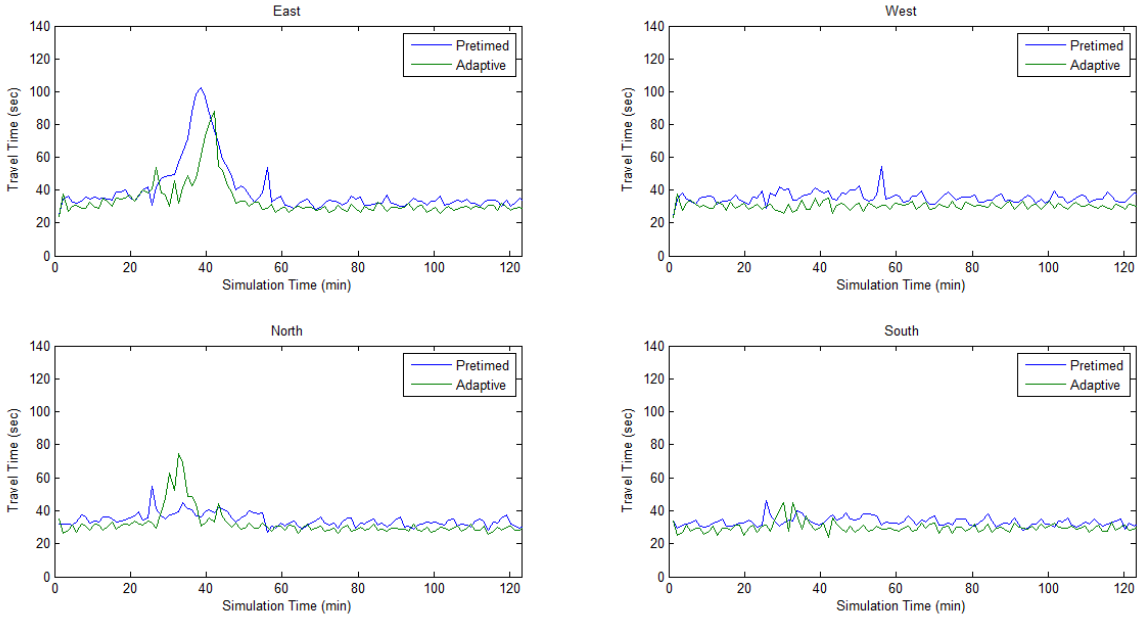


Fig. 8: Travel Time variations over simulation time on all bounds

[6] C. Somm, Z. Yao, R. German, F. Dressler, "On the need for bidirectional coupling of road traffic microsimulation and network simulation" in Proc. International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc), p 41-48, 2008, 9th ACM International Symposium on Mobile Ad Hoc Networking and Computing - Proceeding of the 1st ACM SIGMOBILE Workshop on Mobility Models 2008, Models'08

[7] iTETRIS project official website: http://www.ict-itetris.eu/

[8] MITSIM official website: http://web.mit.edu/its/mitsimlab.html

[9] TraNS official website: http://lca.epfl.ch/projects/trans

[10] Vineet Kumar, Lan Lin, Daniel Krajzewicz, Fatma Hrizi, Oscar Martinez, Javier Gozï£¡lvez, Ramon Bauza, "iTETRIS: Adaptation of ITS Technologies for Large Scale Integrated Simulation", VTC Spring'2010. pp.1~5

[11] Traffic simulator, SUMO, official website: http://sumo.sourceforge.net/

[12] Daniel Krajzewicz, Michael Bonert, and Peter Wagner. "The Open Source Traffic Simulation Package SUMO". In RoboCup 2006 Infrastructure Simulation Competition, Bremen, Germany, 2006

[13] Network Simulator, ns-3, official website: http://www.nsnam.org/

[14] T. R. Henderson, M. Lacage, and G. F. Riley. "Network Simulations with the ns-3 Simulator". Demo paper at ACM SIGCOMM'08, August 2008.

[15] J. Maneros, M. Rondinone, A. Gonzalez, R. Bauza and D. Krajzewicz, "iTETRIS Platform Architecture for the Integration of Cooperative Traffic and Wireless Simulations" in Proc. IEEE Vehicular Technology Conference, 2010.

[16] "Traffic Control Systems Handbook", Federal Highway Administration Report FHWA-HOP-06-006, October 2005, section 1.5: "System Evolution".

[17] B. Zhou, J. Cao, X. Zeng, and H. Wu, "Adaptive Traffic Light Control in Wireless Sensor Network-Based Intelligent Transportation System", in Proc. VTC Fall, 2010, pp.1-5.

[18] Dissanayake, D.T.; Senanayake, S.M.R.; Divarathne, H.K.D.W.M. and Samaranayake, B.G.L.T.; "Real-Time Dynamic Traffic Light Timing Adaptation Algorithm and Simulation Software" in Industrial and Information Systems (ICIIS), 2009 International Conference, Dec 2009, pp.563-567.

[19] Al-Khateeb, K. and Johari, J.A.Y.; "Intelligent Dynamic Traffic Light Sequence Using RFID", in Computer and Communication Engineering, 2008. ICCCE 2008, May 2008, pp.1367-1372.

[20] Patel, M. and Ranganathan, N., "IDUTC: An Intelligent Decision-Making System for Urban Traffic-Control Applications", in Vehicular Technology, IEEE Transactions, VOL. 50, No. 3, May 2001, pp.816-829.

[21] Michel Ferreira , Ricardo Fernandes , Hugo Conceição , Wantanee Viriyasitavat , Ozan K. Tonguz, Self-organized traffic control, Proceedings of the seventh ACM international workshop on VehiculAr InterNETworking, September 24-24, 2010, Chicago, Illinois, USA

[22] V. Gradinescu, C. Gorgorin, R. Diaconescu, V. Cristea and L. Iftode, "Adaptive Traffic Lights Using Car-to-Car Communication", Vehicular Technology Conference, 2007. VTC2007-Spring. IEEE 65th In Vehicular Technology Conference, 2007. VTC2007-Spring. IEEE 65th (2007), pp. 21-25.

[23] Shih-Fen Cheng, Epelman, M.A., Smith, R.L., "CoSIGN: A Parallel Algorithm for Coordinated Traffic Signal Control",VOL. 7, No. 4, Dec 2006, pp.551 - 564.

Innovated Processes of the Imaging Department of a Health Institute Evaluated by Learning Curves

J. García-Porres and M. R. Ortiz-Posadas
Department of Electrical Engineering. Universidad Autónoma Metropolitana-Iztapalapa. D.F. México

Abstract - *The National Institute of Respiratory Diseases is a public hospital located in Mexico City which, in 2007 acquired a PACS-RIS system in its Imaging Department (ID), with the objective to enhance their service. It was necessary to analyze the process that was held in the ID and the related four sub-processes: Reception, X Ray, Computed Tomography, and Diagnosis. In order to make these processes more efficient it was generated a pilot program that consisted in the implementation of 13 innovations related with personnel training, workflow changes, and workload distribution. The objective of this work was to evaluate the impact of the innovations implemented in all the processes, through the estimation of learning curves as an analytical model to evaluate performance in a time frame. It was observed an increase in the performance of the ID general process and in the four sub-processes also. In the case of the information flow, related with data processing in the system, the increase in the learning percentage at the 30 days of the innovations implementation was from 24% to 44%. In the wait times case, the learning increment was 12% for Reception and up to 24% for X Ray and CT.*

Keywords: Learning curves, imaging department, innovated processes, PACS.

1 Introduction

The National Institute of Respiratory Diseases (INER for its Spanish acronym), is a third level public hospital located in Mexico City which, in 2007 acquired a PACS-RIS system in its Imaging Department (ID), with the objective to enhance their service. The introduction of this new technology represented an important change in the way the service is provided. Because of this, it was necessary to analyze the general process that was held in the ID, and the related four sub-processes: Reception, X Ray, Computed Tomography and Diagnosis; with the objective to make these processes more efficient and utilize the available Institute's resources in a more balanced way, there were detected the non-value added activities and the opportunity areas [1]. After this, it was generated a pilot program that consisted in the implementation of 13 innovations divided in three general categories: a) personnel training, b) workflow changes, and c) workload distribution. The objective of this work was to evaluate the impact of the innovations implemented in the four sub-processes and the general ID process, through the estimation of learning curves as an analytical model to evaluate performance in a timeframe. The learning curves serve also as a time prognosis of when the maximum learning will be reached by the actors involved in the activities of each process.

2 Methodology

2.1 Indicators definition

To measure the performance of each process, it was defined a group of variables related with the PACS-RIS information flow, and another group related with wait times related with human factor. With these variables it was defined a group of normalized indicators within a [0, 1] interval, associated with a relevance factor according to the feasibility of the innovation realization and the impact in the process.

2.2 Processes measurement

To measure the time variables, first it was determined a sample of 50 patients calculated from the standardized value of 3500 patients attended in average per month at the ID [2]. The sample was divided in 25 X Ray patients and 25 Computed Tomography patients, randomly selected during 4 days. Then, it was measured the lapse of each patient, since the arrival to the ID until the corresponding study was available in the PACS-RIS, and/or the printed study was delivered. The measurements related with the information flow variables were obtained from the PACS-RIS, analyzing the total registered work per shift during the measurement period. The information was obtained in three different time frames: in 0 days (t_1) to know the performance of the original process (M_1); and after the innovations implementation in t_2=15 days (M_2) and t_3=30 days (M_3).

2.3 Learning curves

A learning curve is a function (1) that represents the performance of someone that learns a new ability as a function of time [3], where N is the maximum learning level. The equation (1) describes an exponential saturated function in which, to obtain the value of the coefficient k, it is necessary to apply a linear transformation, where k represents the straight-line equation [4].

$$P(t) = N - e^{-kt} = 1 - e^{-(mt-b)} \quad (1)$$

After this, it was graphed the learning curve P(t) and in each case it could be forecasted the time in which 80% of the maximum learning level might be reached, evaluating the t variable when P(t)=0.8, this value is considered an adequate learning level considering the human factor in the process performing.

For each process, there were generated two learning curve types: one for information flow and another one for wait times. In both cases the expected function (1) has the maximum learning value in 1 (N=1) on the y axis and is equal to 100%.

3 Results

3.1 Indicators definition

To know the processes performance there were defined 25 variables (V_i) and 19 indicators (I_i) related with the information flow through the PACS-RIS system (Table 1) or wait times (Table 2). It is important to clarify that not all the indicators have the same impact in the process; because of this it was assigned a weigh (P_i) according to its relevance. Note that in Table 1 the indicators C_R, D_R, E_R, II_X, C_{TC}, and I_D have the major relevance (P_i=1) and the rest of the indicators have a value between 0.38 and 0.75. In the case of the wait time indicators (Table 2), three have the relevance (P_i=1) and the two remaining have the minimum relevance (P_i=0.25).

3.2 Processes measurement

The first measurement (M_1) was performed to know the original process performance and the result of the information flow indicators which is showed in Table 1. Note that some indicators have a value of zero. This is because the PACS-RIS at that time was not fully utilized because the Reception personnel didn't program patients in the agenda, the technicians didn't conclude studies, and the radiologists didn't diagnose the studies. In the case of the indicator D_R that is related with the installation of a studies printer in Reception, the value of zero represents that at that time this resource wasn't available. For the time indicators, the M_1 corresponding values (Table 2), represent the initial reference of wait time for the related activities. What it is expected in the subsequent measurements is that the wait times are reduced.

Once the innovations were implemented, there were performed two more measurements, in t=15 days (M_2) and in t=30 days (M_3). The result of the information flow indicators is presented in Table 1. Note that the II_R indicator obtained a value of zero in the three time frames because it was not possible to launch the RIS agenda to program the patients' appointments; however in the rest of the indicators there was an increase in the subsequent measurements, which means that there was an enhancement in the performance of the sub-processes of the ID. In the case of the wait time indicators, the result is shown in Table 2. Note that in general, the wait times decreased, but in the B'$_R$ and C'$_R$ indicators there was no increase probably because the usage of the RIS is related and it was not possible to launch the system's agenda to automate the studies programming.

3.3 Learning curves

Next it is illustrated the procedure to calculate the learning curves, applied to the X Ray sub-process [5].

1. Calculate the indicators change Δ_j for each measurement period using the equation (2).

$$\Delta_j = \frac{\sum_{i=1}^{n}(I_i)(P_i)}{\sum_{i=1}^{n} P_i}, j \rightarrow \{1, \ldots, 3\} \quad (2)$$

The product of the indicator (I_i) by the corresponding weigh (P_i) in each measurement period (M_1, M_2, and M_3), and the sum are shown in Table 3.

Table 3. Product $(I_i)(P_i)$ of the indicators I_X and II_X in the three measurement periods

Ii	**P_i**	**M_1**	**$M_1(P_i)$**	**M_2**	**$M_2(P_i)$**	**M_3**	**$M_3(P_i)$**
I_X	0.75	0.28	0.21	0.60	0.45	0.55	0.41
II_X	1.00	0.00	0.00	0.43	0.43	0.96	0.96
$\sum$	**1.75**		**0.21**		**0.88**		**1.37**

2. Apply the linear transformation LT (3) to the indicators change Δ_j (Table 4), graph t vs LT, and obtain the straight line equation Y(t); meaning the coefficient k of (1). In Graphic 1 it is shown the curve and the straight line equation for each set of indicators type. Note that in the information flow and in wait times, there was an increase in the result of the indicators, which means that the process innovations had a positive impact in the workflow and in the patient by the reduction of wait times.

$$LT = log_{10}\left(\frac{1}{(1-P)}\right) \quad (3)$$

Table 4. Indicators change Δ_j and linear transformation LT for the X Ray sub-process

	Information flow		Wait time	
Time [days]	Δ_j	LT	Δ_j	LT
0	0.118	0.061	0.686	0.503
15	0.494	0.290	0.877	0.910
30	0.785	0.665	0.911	1.051

3. Obtain the equation of the learning curve for information flow $P(t)_{FI}$ and wait times $P(t)_{TE}$ of the X Ray sub-process, using (1) and graph.

$$P(t)_{FI} = 1 - e^{-(0.02t+0.036)} \quad (4)$$

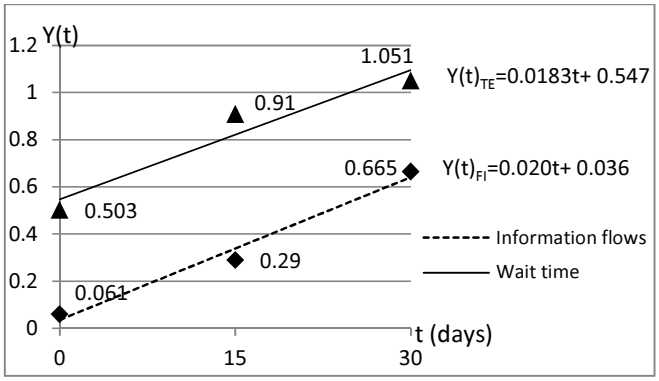
$$P(t)_{WT} = 1 - e^{-(0.02t+0.055)} \quad (5)$$

Table 1. Variables and indicators of information flow in the ID

Process	V_i	Variable description	I_i	P_i	M_1	M_2	M_3
Reception	V_1	Number of applications without error	$B_R=V_1/50$	0.75	0.38	0.86	0.84
	V_2	Number of patients with payment	$C_R=V_2/50$	1.00	0.42	0.94	1
	V_3	Measure of subsequent patient attention	$D_R=V_3/50$	1.00	0.36	0.66	0.7
	V_4	Printer installation in Recepction	$E_R=V_4$	1.00	0	1	1
	V_5	Number of CT studies in the RIS agenda	$I_R=V_5/V_6$	0.63	0	0	0.37
	V_6	Number of CT studies taken per shift					
	V_7	Number of programmed studies in RIS	$II_R=V_7/V_8$	0.63	0	0	0
	V_8	Number of X Ray studies taken per shift					
	V_9	Number of incorrect patient data registers	$III_R=V_9/V_{10}$	0.50	0.51	0.10	0.30
	V_{10}	Number of patients attended in one shift					
	V_{11}	Number of inconsistent registers	$IV_R=V_{11}/V_{10}$	0.38	0.14	0.00	0.25
RX	V_{12}	Number of incorrect registers between CR and RIS	$I_X=V_{12'}/V_{12}$	0.75	0.28	0.60	0.55
	$V_{12'}$	Number of correct registers between CR and RIS					
	V_{13}	Number of X Ray studies concluded in one shift	$II_X=V_{13}/V_{14}$	1.00	0	0.43	0.96
	V_{14}	Number of X Ray studies completed in one shift					
CT	V_5	Number of CT studies in the RIS agenda	$A_{CT}=V_5/V_6$	0.63	0	0	0.04
	V_6	Number of CT studies taken per shift					
	V_{15}	Number of CT studies concluded in one shift	$B_{CT}=V_{15}/V_6$	1.00	0	0.77	0.95
	V_6	Number of CT studies taken in one shift					
Diagnosis	V_{16}	Hiring of a RIS administrator	$A_D=V_{16}$	0.63	0	1	1
	V_{17}	Number of studies diagnosed in PACS					
	V_{18}	Number of studies diagnosed in one shift	$B_D=V_{17}/V_{18}$	1.00	0	0	0.33

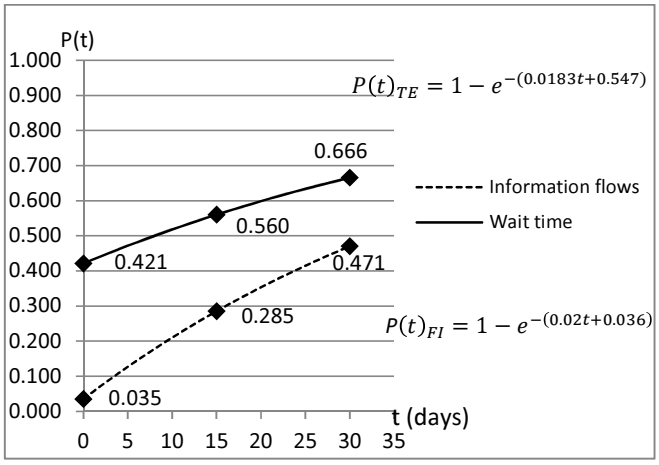
Table 2. Variables and indicators of wait time in the ID

Process	V_i	Variable description	I_i	M_1	M_2	M_3
Reception	V_{19}	Patient's arrival time	$A'_R=V_{20}-V_{19}$	04:01	00:48	00:32
	V_{20}	Patient's attending time				
	V_{20}	Patient's attending time	$B'_R=V_{21}-V_{20}$	24:29	10:16	17:28
	V_{21}	Availability time of data in RIS				
	V_{22}	Patient's calling time	$C'_R=V_{23}-V_{22}$	23:00	16:18	17:28
	V_{23}	Study availability time in RIS				
RX	V_{22}	Patient's calling time	$A'_X=V_{22}-V_{24}$	14:08	05:31	04:34
	V_{24}	Application arrival time to X Ray area				
TC	V_{22}	Patient's calling time	$A'_{CT}=V_{22}-V_{25}$	17:29	17:25	05:16
	V_{25}	Application arrival time to CT area				



Graphic 1. Linear transformation of indicators Δ_j of the X Ray sub-process

The increase in the learning level was clearly reflected in Graphic 2. Note that for the wait times there was reached a 66% at the 30 days measurement, this meant a global increase of 24% between the first measurement M_1 and the third one M_3. For information flow, the behavior was similar, there was reached a 47% of learning level, with an overall increase of 44% during 30 days, this is because the initial learning was practically null.



Graphic 2. Learning curves of the X- Ray sub-process

4. Calculate the time in which it is forecasted that the process will reach 80% of the maximum learning, for information flow and wait times using the equations (4) and (5) respectively.

$$P(t)_{FI} = 1 - e^{-(0.02t+0.036)} = 0.80$$

$$P(t)_{WT} = 1 - e^{-(0.02t+0.055)} = 0.80$$

$$t_{FI} = \frac{ln(0.2)+b}{-m} = \frac{ln(0.2)+0.036}{0.02} \simeq 79\ days\,, \quad t_{WT} \simeq 58\ days$$

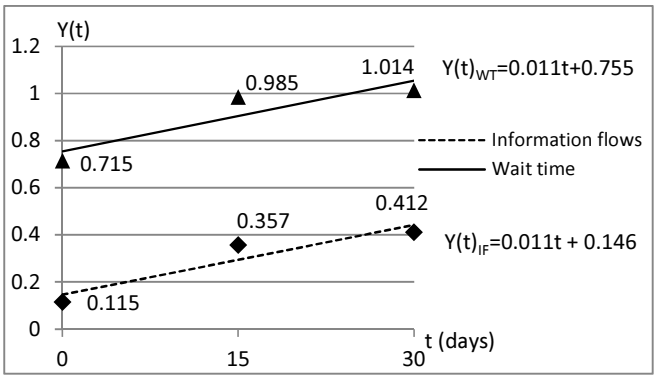
According to the prognosis, the 80% of the learning for information flow will be reached in 79 days, and for the wait times in 58 days.

This same procedure was performed for the other three sub-processes (Reception, Computed Tomography, and Diagnosis). In Table 5 it is shown the change in the indicators Δ_j for each process in each time frame, as well as their linear transformation. The learning curves for each one of the sub-processes are presented later.

Table 5. Indicators change Δ_j and linear transformation LT for the three processes R, TC y D.

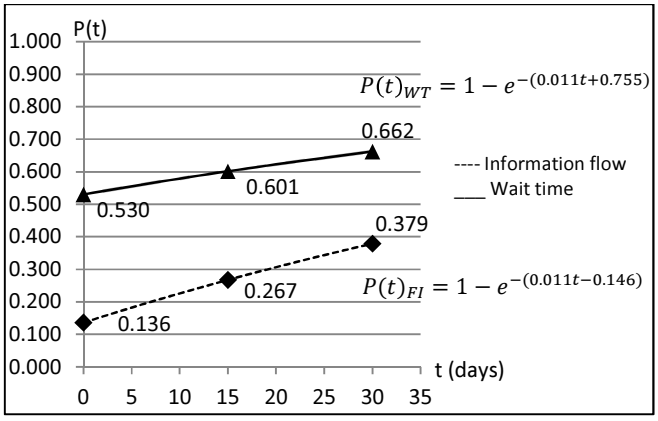
Process	M_i t[days]	Information f Δ_j	LT	Wait time Δ_j	LT
Reception	0	0.233	0.115	0.807	0.715
	15	0.561	0.357	0.896	0.985
	30	0.613	0.412	0.903	1.014
Computed Tomography	0	0.000	0.000	0.783	0.664
	15	0.473	0.278	0.790	0.678
	30	0.597	0.395	0.940	1.222
Diagnosis	0	0.000	0.000		
	15	0.263	0.133		
	30	0.509	0.308		

Learning Curves in Reception. In both information flow and wait times, there was an increase in the result of the indicators (Table 5) and as in the previous case, it was graphed t vs LT (Graphic 3) and it was obtained the straight line equation Y(t) in each case.

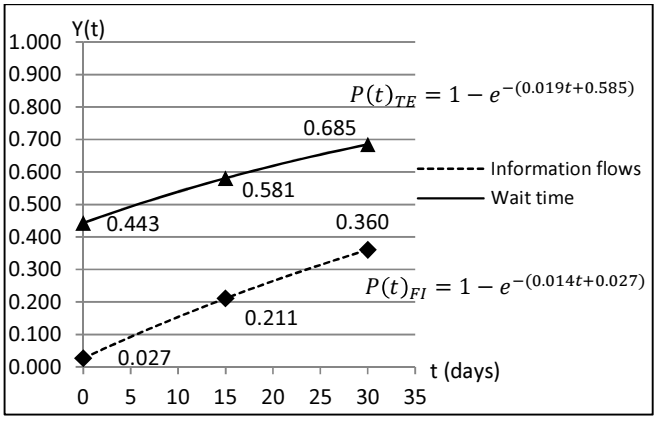


Graphic 3. Linear transformation of indicators Δ_j of the Reception sub-process

In the other hand, the learning curve equation for information flow and wait times was calculated using (1) and are shown in Graphic 4. The increase in the learning levels for the Reception sub-process was reflected clearly in both learning curves. Note that for wait times there was reached a 66% in 30 days, with an overall increase of 12% between the first measurement (M_1) and the third one (M_3). For information flow the behavior was similar, it was reached a 37% of learning level with an overall increase of 24% in 30 days. To calculate the time in which P(t)=0.80, it was obtained that for information flow the 80% is reached in 135 days and in 40 days for wait times.



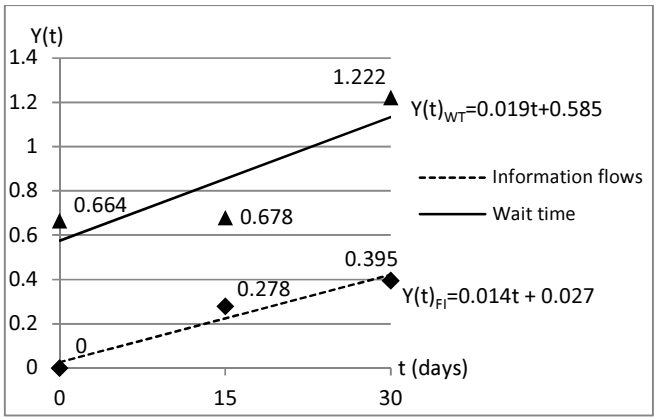
Graphic 4. Learning curves of the Reception sub-process



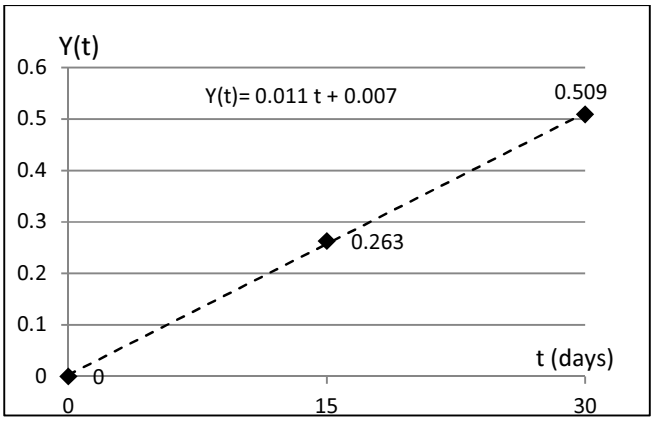
Graphic 6. Learning curves of the Computed Tomography sub-process

Learning Curves for Computed Tomography [6]. In this sub-process there was an increase too in the result of the indicators (Table 5) and in the same way compared to the previous case, it was graphed t vs LT (Graphic 5) and it was obtained the straight line equation Y(t) for each case.

Learning Curve for Diagnosis. In this sub-process there was an increase in the result of the information flow indicators (Table 5), it is important to remember that in this case there were no wait times indicators (Table 2). In the same way it was graphed t vs LT (Graphic 7) and it was obtained the straight line equation Y(t).



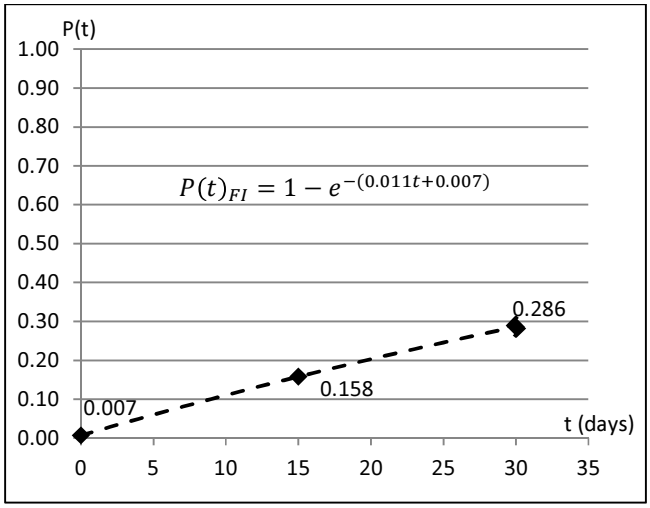
Graphic 5. Linear transformation of indicators Δ_j of the Computed Tomography sub-process



Graphic 7. Indicators linear transformation of the Diagnosis sub-process

The equation of the learning curve for each case was calculated using (1) and it is shown in Graphic 6. The increase in the learning levels for the Computed Tomography sub-process was clearly reflected in the learning curve. Note that for wait times there was reached a 68% in 30 days, with an overall increase of 24% between the first measurement (M_1) and the third one (M_3). For information flow, it was reached a 36% of learning level with an overall increase of 34% because the initial learning level was practically null (2.7%). To calculate the time in which P(t)=0.80, it was obtained that for information flow the 80% is reached in 113 days and in 54 days for wait times.

The learning curve equation is shown in Graphic 8. The increase in the learning levels for this sub-process was minimum. Note in the curve that a 28% was reached at 30 days, with an overall increase of the same value, because the initial learning was very near to zero. This is because the radiologists were not using th system to visualize the images or perform the studies diagnosis. The estimated time to P(t)=0.80 resulted in t=146 days. This time can be decreased probably with a continuous training program on the use and handling of the PACS system, so that the radiologists incorporate these activities to their daily routine activities.



Graphic 8. Learning curve of the Diagnosis sub-process

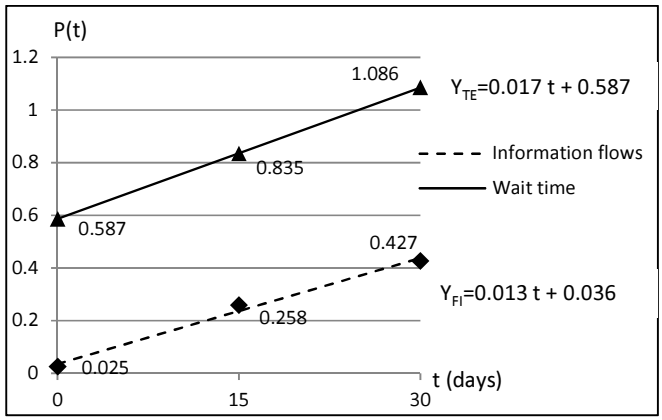
3.4 Global learning curve of the imaging department (ID)

To develop the global learning curve of the ID it was considered the average of the result of the indicators in each measurement period for each sub-process; for both information flow and wait times using equation (2) it was performed the linear transformation (Table 6). It was graphed t vs LT (Graphic 9) and it was obtained the straight line equation $Y(t)_{FI}$ and $Y(t)_{TE}$.

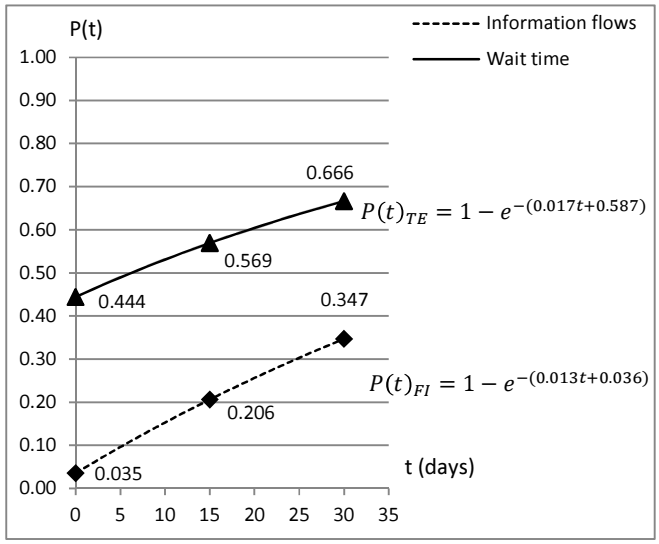
Table 6. Result of the indicators and its average in each measurement period for each sub-process

Days	RX	R	CT	D	Δ_i	LT
Information flow						
0	0.118	0.233	0	0	0.090	0.025
15	0.494	0.561	0.473	0.263	0.448	0.258
30	0.785	0.613	0.597	0.509	0.626	0.427
Wait times						
0	0.686	0.807	0.783	-	0.741	0.587
15	0.877	0.896	0.790	-	0.854	0.835
30	0.911	0.903	0.940	-	0.918	1.086

In the other hand, the equation of each learning curve was calculated using (1) and is shown in Graphic 10. The overall learning increase for the global ID process is clearly shown in both curves. Note that for wait times, a 66% was reached in 30 days with an overall increase of 22%. For information flow, it was reached a 34% learning level with an overall increase of 31%, because the initial learning level was practically null (3.5). The increase in the learning level was due to the correct PACS-RIS usage by the ID personnel, after the training. In the other hand, to calculate the time in which P(t)=0.80, it is forecasted that in 121 days an 80% of learning level will be reached in the global ID process; for wait times there are 60 days with a growing positive tendency in both cases.



Graphic 9. Global linear transformation of indicators of the ID process in each measurement time



Graphic 10. Global learning curves of the Imaging Department process

4 Conclusions

In general, it was observed an increase in the performance of the ID general process and in the four sub-processes as well. In the case of the information flow, related with data processing in the system, it was observed initially that the personnel use of the PACS-RIS system was limited because the indicators demonstrated, in t=0 a practically null value; such as 0.007 for the Diagnosis sub-process (Graphic 8), or 0.136 for the Reception one (Graphic 4). In this sense, it is important to note that the increase in the learning percentage, at the 30 days of the innovations implementation was from 24% to 44%. In the wait time case, the estimated learning initially resulted better, in general superior to 40%, which means that the personnel had a good handling of the

patients within the whole process. This means, a lower learning increment of 12% Reception, up to 24% for X Ray and CT.

It is important to say that the learning curves resulted a good mathematic tool to analyze the impact that the implemented innovations had in all the ID's processes, because the individual learning is the enhancement that is obtained when the people repeat the process and acquire ability, efficiency, or practice from their own experience. Also, the learning level of the organization it is also the result of the practice, that comes from administrative changes, equipment, and product and processes re-design.

5 References

[1] J García-Porres, MR Ortiz-Posadas, and AB Pimentel-Aguilar: *Lean Six Sigma Applied to a Process Innovation in a Mexican Health Institute's Imaging Department*. Proceedings 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Vancouver, Canada, 5125-8: 2008.

[2] National Institute of Respiratory Diseases. *Annual and monthly report generated at the Imaging Department*. May 2008. (*In Spanish*)

[3] Teplitz CJ: The Learning Curve Deskbook: A Reference Guide to Theory, Calculations, and Application. US: Quorum Books. 1991

[4] J Holman: Experimental Methods for Engineers, 7 Ed, US: McGraw Hill, 2000.

[5] García-Flores SE, García-Porres J and Ortiz-Posadas MR: *Learning Curves of the X Ray Innovated Process in the Imaging Department of the National Institute of Respiratory Diseases, Mexico.* Proceedings 32th Annual International IEEE EMBS Conference. Buenos Aires, Argentina, 450-453: 2010.

[6] García-Flores SE y Ortiz-Posadas MR: *Learning Curves of the CT Innovated Process in the Imaging Department of the INER.* Proceedings XXXIII Mexican Congress of Biomedical Engineering, Guanjuato, México, 2010. *(In Spanish)*

A new Approach to Judgments on Video Applications for Required Contents

Mustafa Rashed [1,2,4], Dingju Zhu [1,2,3,*]

[1] Laboratory for Smart Computing and Information Science, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences-518055 Shenzhen, China
[2] Graduate University of Chinese Academy of Sciences-100049 Beijing, China
[3] School of Electronic Engineering and Computer Science, Peking University- 100871 Beijing, China
[4] Department of Computer Science and Engineering, University of Chittagong-4331 Chittagong, Bangladesh

Abstract *Searching desired videos is now challenging due to scattered collection of huge amount of videos all over the world. It is very difficult to find appropriate different video applications for different video contents required by different users from repositories. Traditionally, the selection is based on user's common sense or expert recommends, however different users have different choices and different experts have different recommendations, and only some of the judgments are luckily appropriate, and when the judgments are mistake, the video content required by different users will not be fully met by the selected video application, which will decrease the video service quality and will waste the video resources for not being fully utilized. This paper contributes prioritizing videos by using Analytical Hierarchical Process (AHP) for selecting and ranking particular video. In this research a simple case study presented over some classification of videos where particular object/person need to be identified. It has been shown that best appropriate for selecting particular object/persons from video database is shot-Stock videos, and then is document, research, I-Video, VOD one by one.*

Keywords: AHP, MCDM, eigenvector, pair wise matrix

1 Introduction

Video retrieval system is widely used where movies are stored in a video server [1,8]. Video retrieval system provides a flexible searching environment to the users from a large set of video titles. User can view full or part of a video without interruption. The selection process is mainly based on bibliographic data such as title, release year, genre or directors/actors names. It is impor*tant to selection of appropriate video criteria from large repositories. Traditional Video retrieval service doesn't provide prioritization of video criteria. Some decision support techniques can enhance the selection process and also satisfies prioritization of videos [8]. In this paper Analytical Hierarchical Process (AHP) [3] incorporated for efficient, prioritized and consistent video selection. AHP is one of the ubiquitous approaches to solve Multi-Criteria Decision Problem (MCDP). It is to be noted that, selection of videos is a multi-criteria decision problem since it has wide spread categories. This are basically in two classes: i) content oriented and ii) content less videos. These two categories further form a hierarchy of video classification. Figure 1 depicted trees of video classification.

In this figure there are five alternatives of video applications such as: VOD, I-video, Shot-Stock, research and document. These applications further contain some selection criteria for example bibliographic, compositional, structural, topic-content and sensory content. The video applications and its different criterion are described below [1,7,9].

Video-On-Demand services (VOD): In this video application users can query on-demand videos from server. Bibliographic, structural and topic-content required for selection process.

Interactive video applications (I-Video): In VOD servers videos stored in linear fashion [8]. Interactive video applications provide the access of scenes in any order. For example interactive news on demand service [8] facilitates interactive video application. Users can select news on topic and image content requesting the system in which sequence the news videos should be played. This service requires structural and compositional data of video information.

Shot-stock applications: To find particular objects and persons, shot-stock applications are used. Television directors and reporters typically used this application [10].

*Corresponding author. Address: Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences-518055 Shenzhen,China.E-mail:dj.zhu@siat.ac.cn;Tel.:+86-13316588865

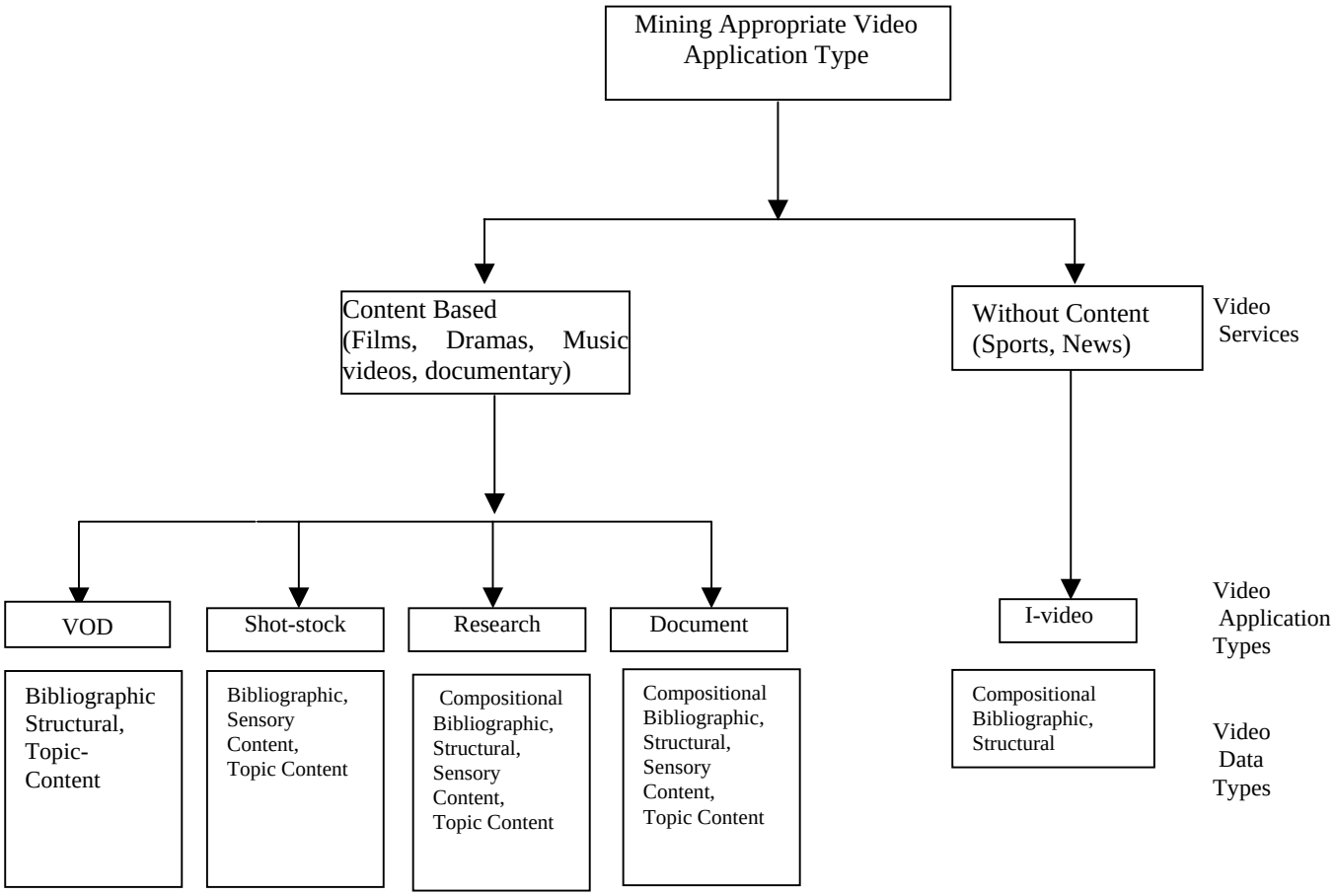


Figure 1: Video Classification Tree

A video shot is a continuous recording of a section of a film. Users select a certain image from shots for their video queries. It requires bibliographic, sensory content and topic content.

Program research applications: Users can collect historical video data for research or personal archives using this video application. "*Peoples Century*" is the biggest historical documentation series taken by BBC [11]. They used this research application of videos for achieving the goal. Video topic data is essential for this kind of research.

Video documentation applications: This sort of applications facilitates users to find appropriate documents of the real world. Anthropology [12], Hand Craft [13] and user requirements analysis [14] are the examples of these applications. Video documentation applications are distinguished to other applications because of using supplementary video information data type.

Compositional Data: Video data are mainly composed in two categories: video documents and media data. Compositional data is acted as a bridge between these two categories. The resultant video has the capability of handling video stream data as well as video documentary data.

Bibliographic Data: This category includes information about the video (title, abstract, subject, and genre) and the individuals involved in the video (producer, director, and actors).
Structural Data: It includes the organization of movie segment such as scene, shot and frames. Traditionally it organized in a hierarchical level. Structural index can also be applicable for audio.

Topic Content: Each video document is based on some context. For example "Olympic 2012" can be used as topic annotation further which can be indexed.

Sensory Content: This is the extractable information from video database. For example content description of content annotations is one of the desired content feature data.

The rest of the paper organized as follows: in section 2 a basic principle of AHP discussed, section 3 illustrates methodology for selection of videos, in section 4 a brief analysis is elucidated for prioritized videos and finally conclusion part is described in section 5.

2. Principles of Analytical Hierarchical Process (AHP)

The basic principle of AHP is based on three steps: i) Problem decomposition ii) Pair wise matrix calculation and iii) summarize the result to find best alternative [14]. Saaty [4] suggested using at least four level of hierarchy in AHP to satisfy the goal. The first level is the goal selection level, second and third are the categories and sub categories and level four is the alternatives. Figure 2 shown a sample AHP structure for video selection. The structure essentially signifies the three level hierarchy structure that indicates the relationship of the goal, the criteria compositional, bibliographic, structural, sensory content, topic content and the alternatives (category A, B, C, D and E). Here each category contains some or all properties of video information like compositional, bibliographic, structural, sensory and topic content. The AHP then requires the decision makers to carry out simple pair-wise comparison judgments (illustrated end of this section). It concerns the relative importance of each criterion and specifying a preference for each decision alternative. In this paper the following calculation methods and techniques used for selecting and prioritizing videos. Let *n* video elements to be compared, *C1 ... Cn* denote the relative weight, relevance or significance of *Ci* with respect to *Cj* by *aij* and form a square matrix *A=(aij)* of order *n* with the constraints that *aij = 1/aji, for i ≠ j,* and *aii = 1*, all *i*. Such a matrix is said to be a reciprocal matrix. The weights of the video elements are consistent if *aik = aijajk* (transitive) for all i, j and k. The next step is to find a vector ω of order n such that $A\omega = \lambda\omega$. ω is said to be eigenvector of order n and λ is a corresponding eigenvalue. Finally a consistency Index (CI) is to be calculated by the following formula:

$$CI = (\lambda_x - n)/(n-1) \quad (1)$$

It tells that the judgment is consistent and Saaty [4] calculated a large number of random matrices of ascending order and CI of those matrices. Now it need to calculate Consistence Ratio (CR) by dividing CI for the set of judgments by the index for the corresponding Random matrix. According to Saaty [4,5,6] CI<0.1 unless the judgment is inconsistence. If CR =0 then it is to be perfectly consistent.

Pair wise comparison judgments and their numeral values (1-9) between A and B [4]

A is equally important than B →1
A is equal to moderately important then B→2
A is moderately important than B→3
A is moderately to strongly important then B →4
A is strongly important then B→5
A is strongly to very strongly important then B→6
A is very strongly important then B→7
A is very strongly to extremely important then B→8
A is extremely strong important then B→9

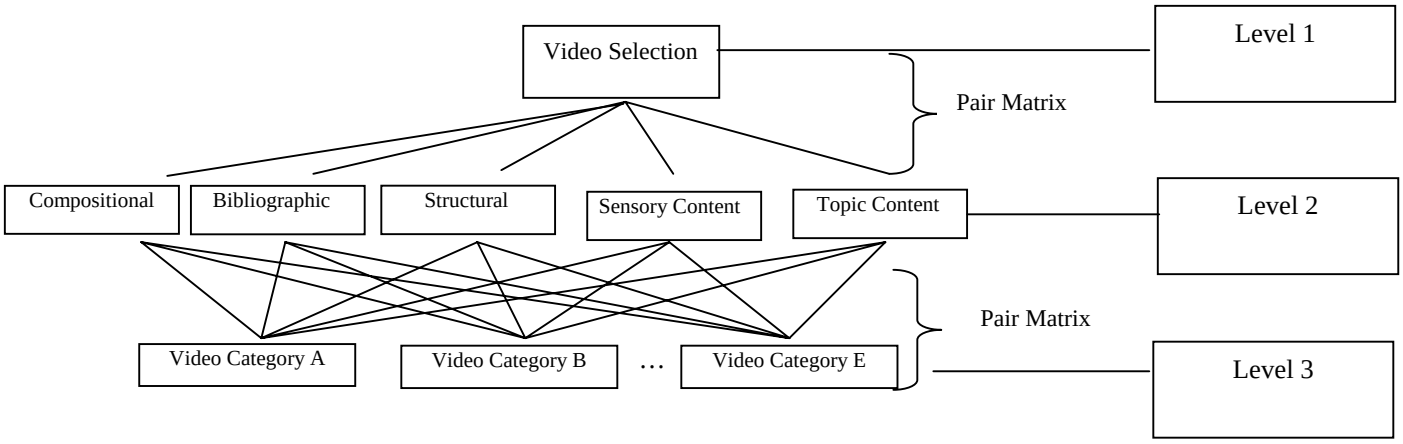


Figure 2: AHP Hierarchical Structure for Video Selection

3. Methodology

To select a desired video the following AHP based approach presented in this paper. The goal is set for Shot-Stock video compatibility in comparison with other available video application types. The first step of this method is to decompose the problems into sub objectives. This phenomenon is presented in figure 2. Here level 1 signifies the overall objective or the focus of the problem, which should always be on the top. Level 2 characterize the criteria of the problem and level 3 represents the alternative that need to be evaluated by the criteria.

AHP Based video compatibility measurement algorithm [4]:

Step 1: Decompose goal into sub objectives.
Step 2: Construction of a pair wise comparison matrix based on expert choices.
Step 3: Calculating eigenvectors
Step 4: Analyze judgment (CI and CR)

Step 5: Item wise comparison matrix construction for best alternatives
Step 6: Calculate eigenvectors for step 5
Step7: Decision judgment

Analytical results of this algorithm illustrated in the following tables 1-4. Numerical values for pair wise judgments depending on expert choices have been constructed based on domain knowledge and different literature of video classification [1], and are verified by consistency test in section 3.1. This values range from 1-9 where 1 denote equal importance and 9 denotes the highest degree of favoritism. This values arranged in a 5X5 matrix A =[a_{ij}] where $a_{ij} = 1/a_{ji}$ which is illustrated in section 2. Main diagonal of this matrix is always one. The reason for this is that the importance between self-criteria is always one. From table 1 it is noticed that compositional video information is five times important than sensory content type video information with respect to the consideration of bibliographic, sensory content and topic content for shot-stock video applications [1]. The reciprocal is as illustrated before. Table 2 is generalized matrix form of table 1. In table 3 eigenvectors [4] are calculated from table 2. The calculation method is as follows:

There are three steps to be followed in deriving eigenvectors for achieving priority vector. The first step is to multiply every value in each row of the pair wise comparison matrix and power the values by 1⁄n (number of dimension) to obtain the total row. In deriving the priority vector, the total row then is divided by the sum of all the total rows. The priority vector is the normalized vector derived after the process is completed.

Table 1: Pair wise comparison matrix

Selection Decision	Compositional	Bibliographic	Structural	Sensory Content	Topic Content
Compositional	1	1	1	5	4
Bibliographic	1	1	1	6	5
Structural	1	1	1	4	3
Sensory Content	1/5	1/6	1/4	1	2
Topic Content	1/4	1/5	1/3	1/2	1

Table 2: Generalized Matrix Form

Selection Decision	Compositional	Bibliographic	Structural	Sensory Content	Topic Content
Compositional	1	1	1	5	4
Bibliographic	1	1	1	6	5
Structural	1	1	1	4	3
Sensory Content	0.2	0.17	0.25	1	2
Topic Content	0.25	0.2	0.33	0.5	1

Table 3: Calculating Eigenvector

Selection Decision	Compositional	Bibliographic	Structural	Sensory Content	Topic Content	N^{th} root	Eigenvector
Compositional	1	1	1	5	4	1.820564	0.290619
Bibliographic	1	1	1	6	5	1.97435	0.315169
Structural	1	1	1	4	3	1.643752	0.262395
Sensory Content	0.2	0.17	0.25	1	2	0.44268	0.070666
Topic Content	0.25	0.2	0.33	0.5	1	0.383081	0.061152
Sum						6.264428	1.00

Calculating Eigenvectors:
Multiply each element in every row and then power of 1/n (N^{th} root in table 4).

Step1: Multiply each element in every row and then power of 1/n.
$(1 \times 1 \times 1 \times 5 \times 4)^{1/5} = 1.820564$
$(1x \ 1 \times 1 \times 6 \times 5)^{1/5} = 1.97435$
$(1x \ 1 \times 1 \times 4 \times 3)^{1/5} = 1.643752$
$(0.2 \times 0.17 \times 0.25 \times 1 \times 2)^{1/5} = 0.44268$
$(0.25 \times 0.2 \times 0.33 \times 0.5 \times 1)^{1/5} = 0.383081$
Step 2: Sum all the rows

1.820564 + 1.97435 + 1.643752 + 0.44268 + 0.383081 = 6.264428

Step 3: Normalize each total of the row by dividing the total row by the total sum of the rows.
1.820564/6.264428 = 0.290619
1.97435/6.264428 = 0.315169
1.643752/6.264428 = 0.262395
0.44268/6.264428 = 0.070666
0.383081/6.264428 = 0.061152

Eigenvector = [0.290619 (2), 0.315169 (1), 0.262395 (3), 0.070666 (4), 0.061152 (5)]
It is important to be noted that, second value has the highest priority. In our example Shot-Stock videos has the highest priority. But later we can investigate it is not the best selection criteria.

3.1 Analyze Judgments

Saaty [8] presented a brief research on AHP and its judgment measures. According to his analysis any candidate item set must have the following properties in order to achieve consistency test.

1. λ_{max} (maximum eigenvector) >= n (number of judgment elements)
2. CR (Consistency ratio) <=0.1 [CR = CI (Consistency Index)/Saaty's judgment number) [4]

Calculating λ_x (maximum eigenvector):
First multiply on the right the matrix of judgments by the eigenvector, obtaining a new vector. The calculation for the first row in the matrix is:

$$1 \text{ X } 0.290619 + 1 \text{ X } 0.315169 + 1 \text{ X } 0.262395 + 5 \text{ X } 0.070666 + 4 \text{ X } 0.061152 = 1.466118$$

and the remaining three rows give 1.597935, 1.334301, 0.37027 and 0.318763. This vector of five elements (1.466118, 1.597935, 1.334301, 0.37027 and 0.318763) is, of course, the product $A\omega$ and the AHP theory says that $A\omega=\lambda max\omega$ so we can now get five estimates of λmax by the simple expedient of dividing each component of (1.466118, 1.597935, 1.334301, 0.37027 and 0.318763) by the corresponding eigenvector element. This gives 1.466118/0.290619 = 5.045 together with 5.070099, 5.085092, 5.239752 and 5.212656. The mean of these values is 5.130481 and that is our estimate for λ_{max}.

In this research maximum eigenvector is 5.130481 [Table 3], consistency index: $(\lambda_{max}-n)/(n-1)$ =0.0326 and consistency ratio (CR) = 0.03262/1.12=0.029125. From this analysis it is clearly shown that the judgment of this research is consistent since CR <0.1.

3.2 Item wise comparison matrix

The previous section deduced eigenvectors of sub problems. This indicates the importance of particular video criterion. For example bibliographic video criteria is suitable in this perspectives, which is depicted in table 3. But it doesn't optimize the alternative video selection. In this research the goal is to prioritize video application for selecting particular object/person, for this reason item wise eigenvector generation is necessary and it is illustrated in the following table 4.

Table 4: Compositional

Selection Decision	VOD	Shot-Stock	Research	Document	I-Video	Eigenvector
VOD	1	1/3	1/2	1	1/3	0.085662
Shot-Stock	3	1	4	5	6	0.495565
Research	2	1/4	1	3	4	0.21851
Document	1	1/5	1/3	1	1/4	0.06733
I-video	3	1/6	1/4	4	1	0.132933

The same approach is applied for other video data types and corresponding Eigenvectors can be generated as per the following illustration:

Bibliographic {(VOD, 0.086219), (Shot-Stock, 0.49879), (Research, 0.070861), (Document, 0.210332), (I-Video, 0.133798)}

Structural {(VOD, 0.086808), (Shot-Stock, 0.502196), (Research, 0.081954), (Document, 0.211768), (I-Video, 0.117274)}

Sensory Content {(VOD, 0.084824), (Shot-Stock, 0.073844), (Research, 0.452498), (Document, 0.219185), (I-Video, 0.169649)}

Topic Content {(VOD, 0.422315), (Shot-Stock, 0.245707), (Research, 0.169505), (Document, 0.06591), (I-Video, 0.096563)}

4. Results

The objective of this research is to prioritize videos according to their criteria. From table 4 it is observed that bibliographic information type obtained the highest rank. It doesn't confirm which type of video applications has the best priority. Table 4 illustrated item wise comparison matrix of videos. It provides judgment weights of each video type.

4.1 Decision judgment

In figure 1, three levels of video classification hierarchy presented. Top level is the goal, the next higher level is criteria and the weights obtained from table 3, finally level three denotes the alternative video selection and judgment weights from table 3-4. Using the following matrix calculation final video selection and judgments can be obtained. In this calculation first matrix contains judgment weights of video type for each criteria and second matrix contains judgment weights of each video criterion. The result shown shot-stock is the highest prioritized video types.

VOD	Shot-Stock	Research	Document	I-Video			
0.085662	0.495565	0.21851	0.06733	0.132933	*	0.290619	Compositional
0.086219	0.49879	0.070861	0.210332	0.133798		0.315169	Bibliographic
0.086808	0.502196	0.081954	0.211768	0.117274		0.262395	Research
0.084824	0.073844	0.452498	0.219185	0.169649		0.070666	Document
0.422315	0.245707	0.169505	0.06591	0.096563		0.061152	I-Video

=	0.106666126 VOD
	0.4532412 *Shot-Stock*
	0.14968246 Research
	0.160944 Document
	0.129467 I-Video

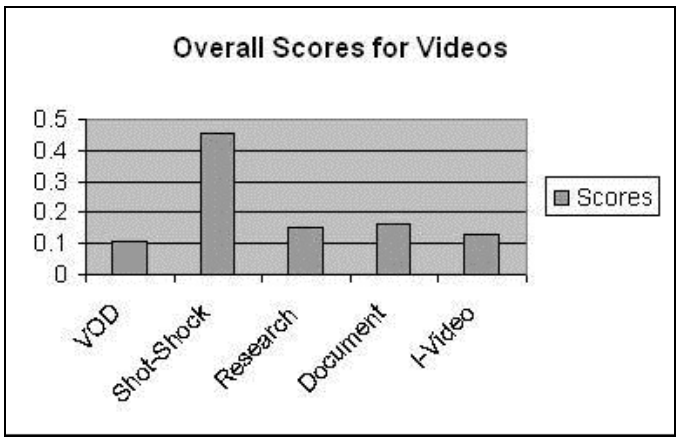


Figure 3: Overall Scores of Videos

5. Conclusion

Widespread large amount of video reveals a proper judgment based selection system. This research presented an AHP based model to select and prioritize video content accurately. It allows a good decision making scenario. Traditional system can't measure the importance of video content and its selection criteria. This paper demonstrated selection judgment of particular object/person from videos. It is shown that the most appropriate type is Shot-Stock, and then is document, research, I-Video, VOD one by one. The result is consistent with expert knowledge. To find particular objects and persons, shot-stock applications are used [10], however obtaining the result through AHP be used to verify the existed expert knowledge is more convincing, and similarly can also be used to create new knowledge such as selection judgment of other things except existing knowledge such as the second, third, fourth, fifth appropriate type for judgment of particular object/person being document, research, I-Video, VOD and selection judgment of particular object/person [10] and selection judgment of interactive news[8] from videos. In addition the initial weights used in this paper are based on some expert choices and wrong placement of initial weights misleads the results, the system robustness can be promoted if a knowledge base inferred the pair-wise judgment matrix.

Acknowledgements
This research was supported in part by National Natural Science Foundation of China (grant no. 61105133) and Shenzhen public technical service platform (grant no. CXC201005260003A).

6. References

[1] Rune Hjelsvold, Roger Midtstraum and Olav Sandsta: "Searching and Browsing A Shared Video Databases", http://home.online.no/~olmsan/publications/papers/mmdbms-chapter.pdf

[2] T. Asahi, D. Turo, B. Shneiderman, Using treemaps to visualize the analytic hierarchy process, Information Systems Research 6(4), pp. 357-375, 1995.

[3] Thomas L. Saaty: "Decision Making with Analytical Hierarchical Process", *Int. J. Services* Sciences, Vol. 1, No. 1, 2008.

[4] Saaty, T.L : "Theory and Applications of theAnalytic Network Process", Pittsburgh, PA : RWS Publications, 2005.

[5] Aryati Bakri, Mahadi Bahari, Azizah Abdul Rahman, Mohd Yazid Pathani : "Review Prioritization Methods in Analytic Hierarchy Process (AHP) ", *Jurnal Teknologi Maklumat,* 2004.

[6] Saaty, T.L : "Decision-making with the AHP: Why is the principal eigenvector necessary", *ISAHP 2001, Berne, Switzerland,* August 2-4, 2001.

[7] J. R. Smith, S-F. Chang, "VisualSEEk: A Fully Automated Content-Based Image Query System", ACM Multimedia Conference, Boston, MA, November, 1996.

[8] T.D.C. Little and D. Venkatesh : " Prospects for Interactive Video-on-Demand", IEEE Multimedia, 1(3):14{24, Fall 1994.

[9] G. Miller, G. Baber, and M. Gilliland : "News On-Demand for Multimedia Networks", In Proceedings of ACM Multimedia 93, pages 383-392, Anaheim, CA, August 1993.

[10] Yi Ding and Guoliang Fan : "Multi-channel Segmental Hidden Markov Models for Sports Video Mining", 2008.

[11] Michael Fleischman, Phillip Decamp & Deb Roy : "Mining Temporal Patterns of Movement for Video Content Classification" , 2006.

[12] Yuya Matsuo, Miki Amano & Kuniaki Uehara : "Mining Video Editing Rules in Video Streams" Multimedia Proceedings of the 10th ACM international conference on Multimedia, 2002.

[13] Sébastien Poullot, Michel Crucianu & Olivier Buisson : "Scalable Mining of Large Video Databases Using Copy Detection", Proceeding of the 16th ACM international conference on Multimedia, ACM Press, PP: 61-70, 2008.

[14] Kumar, N.V. and Ganesh, L.S: "A simulation-based evaluation of the approximate and the exact eigenvector methods employed in AHP", European Journal of Operational Research, 656-662, 1996.

Real-Time Bond Between Social Space Behavior and Spatial Spaces

Ali Jabbari Jahromi[1] , **Azadeh Alighanbari**[2] , **Maryam Shokrollahi**[3]
[1]Architecture Department, Islamic Azad University of Shiraz, Iran
[2] Architecture Department, Islamic Azad University of Shiraz, Iran
[3] E Learning Department, School of Electrical Engineering and Compute,Shiraz University, Iran

Abstract- *Humans and the concept of layers were never separable. In fact human cannot be defined anything but layers. Layers which are always ready to generate new versions of themselves, versions that may have the least similarity to their origin. Just like ever human's irresponsibility and self-imposed ignorance have made him use titles as an effortless approach for reorganization, clarification and even confirmation of these layers. Even in this phase we observe the endless productive hierarchy of titles.*

As philosophers, physiologists, sociologists and cognoscenti's were on an attempt to distinguish human's intellectual layers; medicines, physiologists, engineers and architects dedicated themselves to exploring structural parts of what one can name as human or human related phenomenon's such as society, social behavior , etc.

The fact that mostly assumes obvious; is the necessity of considering a moderate level of awareness between the parallel, complex, none-separable layers to present a general understanding of that concept without involving countless details of each one. This approach is supposed to create harmony between these two different types of layers to fulfill essential needs of each layer in corresponding to the others.

What this article is trying to offer is definition of a dynamic method to reach that mentioned level of understanding to provide the ability of bonding these two types of layers, so any observed change in each would cause an effect to the other. As for this particular case study the goal is to define a bond between urban/architectural space as the most external sample of man's solid layer in logical respond to its physical and intellectual requirements.

Keywords: Urban Space, Practical Chaos

Theory, Fractal, Social Space, Behavioral Science

1 Introduction-Logical Justification

What is going to be discussed here is not an architecture formed object in concrete and stone, but is a flexible tool for understanding of basic architectural elements' strange attractors.

By stating "strange attractors" of architectural elements we attempt to enlighten the characteristic essence of the basic architectural elements in a real-time urban laboratory .As the apparatus, here, is defined by urban/architectural values, the experimental materials would be social behaviors of people and what is to be studied here is the feedback of people's behavior in confronting to those values.

The defined model should dramatize the same chaotic process that people themselves are dealing with without even realizing it's free/oppressed nature.

The new tool has been established to challenge the traditional process; emphasizing the ever-changing mutation of spatial space in the process of logical responding to its vital and basis requirements.

In traditional process of design, object as the final product process is incapable of fulfilling the purpose of its creation. That is due to its incapability of adaptation and compared to the purpose of its creation, "social space behavior" mutation.

The enormous growing speed of mutation and change in social space behavior and chaotic pattern of that behavior accelerate the process of

transforming the final product of design process into spoiled block of urban/ architectural environment called urban hole.

In opposition to that mentioned process, new one at its first step is focusing on the process itself not the final product. Here the process of design is object of design. And at second step it again changes the hierarchy of process and focuses on the cognition process toward design process.

Chaos theory has been used as proper tool for reorganization and cognition of the social space behavior essence. What eventually would be considered as final product is not designer personal subjectivity but is a logical response to those parameters that has necessitated its existence. That would not be the final and precise answer either. But it would represent itself as a continuous, logical alternative, along with the texture's chaotic process.

2 Methodologies

2.1 Basis and concepts

The volition of connecting and joining the necessity of the recognition of the spatial space of changes and the definition of epistemology leads into the production of the capability of proper prediction of the subject related to the social space behavior obvious. Difference in the interpretation of recognition and prediction will clarify itself where the term of prediction is defined as happening and the occurrence of what is exactly expected to. On the other way, as it is used in this research, the term of prediction is none separable from behavioral probabilities of a system.

Intuitive study of social space behavior provides a way for chaos theory to role as a proper and flexible template for interpretation of that. The difference between the meanings of "recognition" and "anticipation" will be clarified in this discussion. The meaning of "anticipation" normally considered as: happening of what is expected to. In the other hand in this paper, "anticipation" would be used as definition of the behavioral probabilities of a system (social space behavior). Intuitive study of social space behavior has made it considerable to interpret it as the phase of the chaos theory.

2.1.1 Defining the chaotic nature of the social space behavior

In the study of social space behavior, tracking the structural signs of a chaotic system will obviate its chaotic nature. Social space shapes its soft/intellectual environment by creation and recreation of symbols, definition of values and taboos and forming signs and social conceptions. It also uses collective awareness and by forming the existing potential interpretive layers of itself to realize its solid surroundings after the interpretation. In other words, in the social space, the environment itself is a part of its entity. According to this speculation, each observed chaoticity in the social space behavior, definitely will have an inside and not an outside source which is a mark to the chaotic nature of social behavior. Another pattern to be observed in the interpretation of social space behavior as a chaotic system is the sensitivity to initial conditions which comes into discussion as we analyze the basis of phenomenones such as revolutions, wars and cognitive evolutions, in such a manner that the butterfly effect firmly reveals its magnificent visage.

On the other hand, stability can be discussed as an unquestioned factor of social space behavior. Although social behavior has survived through huge evolutions in junctures and sometimes in tensions such as these it has been forced to take up formal mutations but in all situations the stability as a result of it's self-modification was a certain and obvious matter. In fact, one of the factors in the stability of the physiological systems is their chaoticity and the placement of their vital variables in strange attractors. In other words, the exit of one variable which is caught in one stranger attractor needs such a change that will change the nature of chaotic system. In other words, from topological point of view, the phase space related to that chaotic system will be affected by such changes and mutation that the phenomenon will transform to another in the process.

It is important to notice that stability of a chaotic system does not put the essential condition such as sensitive dependence to initial conditions, into question but determines the unpredictable behavior of the system in the specific mentioned duration. Rather than that, sensitivity to initial conditions indicates different positions of a chaotic system due to different points of observations in time. In other word, it indicates the position and situation of critical variables of that system in the phase space related to that.

2.1.2 Determination of data and dependent mathematics

Mathematics could be used as a meta-theory in interpretation of data presented by social space behavior and generates them to be suitable for the demonstration of strange attractors which are true represent of the topological nature of the social space behavior. These mathematical functions should be synchronous by the chaotic nature of the social space behavior and the nature of the discussed functions should be in a way that the process models the same chaotic path that they have passed in the upper abstraction, the social space behavior. Fractal mathematics possesses these kinds of abilities and foresight.

Nature of inputs to be used in these functions is so essential to the subject that a proper definition of their characteristic will symbolize them as formal represents of social space behavior and will synchronize to the reality of the nature of the system. Internal data are naturally related to the social space. These data should actually be quantitative to be analyzed, while necessary variables of social space behavior are actually qualitative.

Since in complex systems behavior of a part would cause similar rate of affects on the other parts we expect analyzing a less complicated parameter in such systems would demonstrate a general form of that system. Here by selecting a new variable, we are one step away from the nature of original ones. This kind of distancing away while opens a path to reach a measurable variable, but it is so expectable for the related variables not to demonstrate an exact equivalent behavior because of the chaotic nature of the system. Or in other words, their behavioral phase space or strange attractors may not be equal. In fact, a type of moderation and balance should be considered between finding a proper and practical variable and expecting egregious mistakes of failing in that search.

We assume that variables which have been chosen as represents of social space behavior have definitely become so far away from their original nature that can not exactly represent the original nature of the strange attractions of their clients anymore.

What we hope in, is the definition of the parameter as the input of a superior function in the hierarchy which we believe will become most close to the nature of its client in transformational process of that function. This could be also accurate for the selection of the form generator mathematical interpretive and reproductive equations which demonstrate themselves as the informational data cells.

Therefore, the data as the basis for decision making are assumed to be the outcomes of a chaotic system. Here, chaoticity is interpreted as consistency and validity of the system in just a period of time which will loose those values by the pass of time. Therefore decisions had no more the prospective backwashes and their longitude based on the chaotic nature of the system will lead into the primary nonlinear chain of partial mistakes and eventually will lead into the formation of crisis in the form of a fundamental change and will cause the destruction of the chaotic system. In another form, decision making should form by the perception of the chaotic behavioral process of the system as the nonlinear chain that short-term results of the decision, design and decision making inputs of one design will form. Here, the decision making process, design, in each type and each shape will shape a chaotic process by the assignment to its chaotic nature. Instead by understanding the chaotic process of system's behavior, decision-making should be formed as a none-linear chain of decisions in which short-term outcomes of a decision will be considered as input of another decision and will role as a part of its database. Therefore, any kind of decision making and design by assignment to their chaotic nature will mode a new chaotic process.

3 Practical implementation steps

Interpretation and recreation of the chaotic process which governs the social space behavior logically must lead into the ability of creation of appropriate physical environment- urban/ architecture- which naturally depends on it.

Considering the effect that numerous variables impose on the social space behavior, simplification and reduction will be regarded as the basis of the modeling that should represent the social space behavior.

Final goal is establishing a laboratory for monitoring, dividing and amplifying an understandable and analyzable variable within social space behavior and illustrating that variable in an urban/ architectural context. Eventually it should lead us into understanding a new generation of urban/ architectural nonstandard and noncommittal variables.

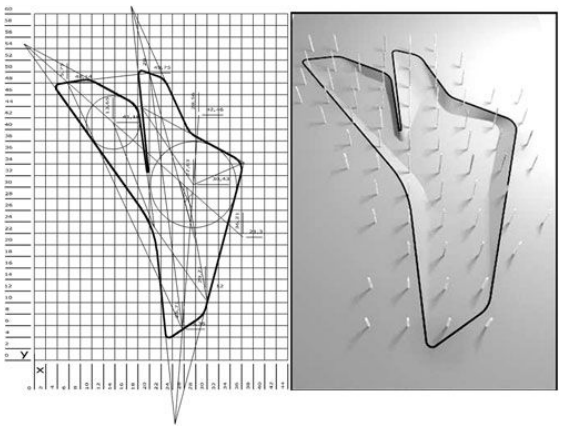


Figure 1- Urban Laboratory

The above mentioned laboratory must role as a part of the urban/ architectural context and be none-separable from that to make uncertainty principle effect the least on this understanding .

In a part of the urban square we have defined a kind of spatial wrinkle for creating the analysis compass and it will be used as a floor for realization of a simple parameter as represent of the social space behavior _ like passing through this urban plot.

This urban plot is designed in a way that includes intuitional urban attractions.

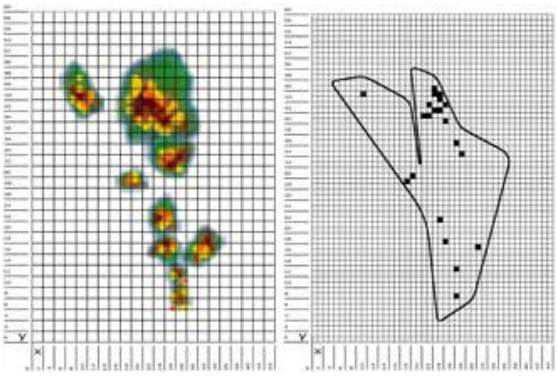


Figure 2- Urban Site Monitoring

What will happen is to monitor, understand and interpret the mentioned parameters in this particular space, and, also generate it in the theoretical framework of chaos theory by using fractal geometrical functions with a software designed for short-term foresight of this parameters from the social space behavior that eventually joins to the final architectural effect which is providing the shade that confines what have been mentioned in foresight and contain more probability for this particular behavior.

3.1 Input

In order to detect the dominant behavior which has been occurred in pass way, the infrared image analyzing method has been used to make snapshots from the environment. Therefore, it is possible to distinguish the density of the population by the assistance of different spectrums produced by the existence of people in the environment. Every five seconds, the system will take photos of the environment which express the existence of the people. In order to net match, every 10 minutes, one hundred and twenty photos produced are entered into the program. For netting match from photos, surroundings and areas of the photos which have less heat are omitted from photos. Then, they are added to each other to reach the hot (agent) points. The output is a matrix of 62 to 44. This matrix is in congruous to the environment of the pass way and each cell will be the representative of one square meter of the area. In every cell of the matrix exists digits such as 0, 1 or 2. Cells with 0 show the nonexistence of people and cells which contain digits except zero show the existence of people. Therefore by the increase in the amount of digits the probability of the existence of the people will be more.

3.2 Process

In this phase, the process of cell expanding is going to be performed till to reach the best quality of the matrix. The results achieved by the expansion will be representative of more density in the existence of the people. In other words, it can be concluded that based on people's behavior more paths will be used, so shadows will be made dynamically in these locations.

In each time of expansion, a new generation of the points will be achieved which will be added to the premier matrix to be used as new agent cells in former expansion. Different phases which are performed in each generation phase are divided into total navigation of the path and production of new images for each non-zero cell which is accomplished by the assignment of digit 1 to new positions. At last, new digits are added by the digits of its corresponding matrix and the result will be production of n^{th} generation.

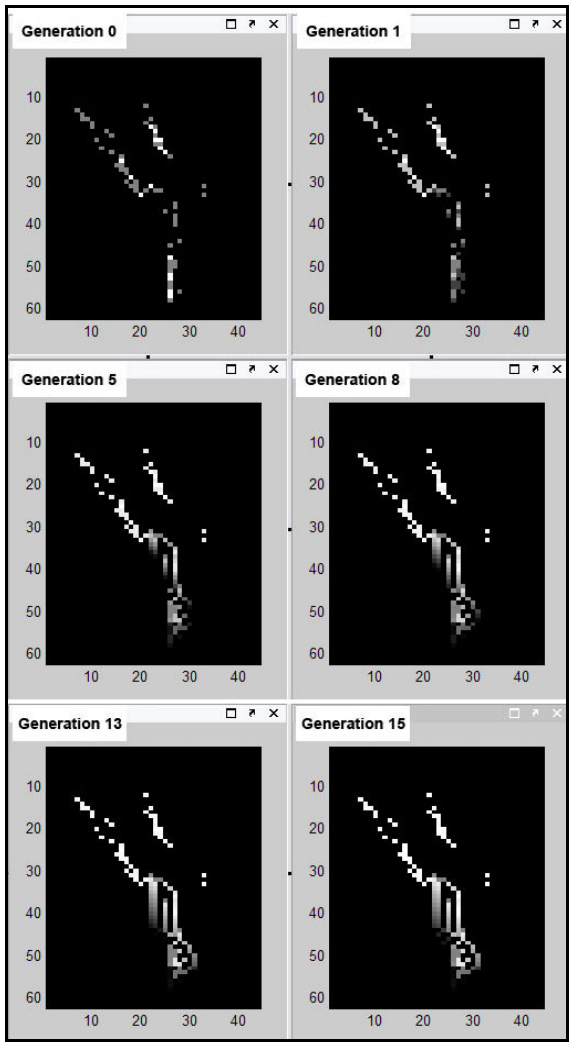


Figure 3- Data Generation Series

The geometric chaotic function which used in the algorithm is a set of none-linear functions. These

functions are based on the behavior of the people in urban fields, achieved based on the defined values which are used as urban designing values. In this geometric function, five critical points are assumed with constant value in pass way area which is as follow: Centers of the fields, centers of the entrance and existence. These points are regarded as tensional points for places which are attraction points (figure for the 5 coordination). If it is supposed that the new coordination is (X_2, Y2), then we will have:

$$X_2 = (A_1 / d_1^2) + (A_2 / d_2^3) + (A_3 / d_3^3) + (A_4 / d_4^2) + (A_5 / d_5^2) + X_1$$

$$Y_2 = (B_1 / d_1^2) + (B_2 / d_2^3) + (B_3 / d_3^3) + (B_4 / d_4^2) + (B_5 / d_5^2) + Y_1$$

In which, (A_1,B_1) ، (A_2,B_2)، (A_3,B_3) ، (A_4,B_4) و (A_5,B_5) are critical points and d is representative of the distance of the agent (X_1,Y_1) from each critical points. This geometric function is set, based on the distance of the agent points in proration to the critical centers.

```
%% ESTIMATE PEOPLE BEHAVIOR IN PASS WAY
%% PART 1 : INPUT
% Number of picture for analyzing
contPic = 120;
% Pass way & picture dimensions
picSize=[62,44] ;
% Set threshold to ommite margines
threshold = 0.96;
picTemp =0;
picRes = picTemp;

for(i=1:contPic)
    % Input pictures for neting match
    s1 = 'infraPics\';s2 = num2str(i); s3 =
'.jpg';
    str = [s1 '' s2 '' s3];
    picTemp=imread(str);
    picTemp=imresize(picTemp,picSize);
    picTemp = im2bw(picTemp,threshold);
    % Matrix 'picRes' contain result of net
match
    picRes=picRes+picTemp;
end
[m,n] = size(picRes);
gTemp = zeros(picSize);
%% PART 2 : PROCESS
% 5 Constant critical points
```

```
A1 = 5.36 ; B1 = 25.71 ;
A2 = 30.43 ; B2 = 27.63 ;
A3 = 41.18 ; B3 = 13.65 ;
A4 = 48.14 ; B4 = 5.75 ;
A5 = 49.75 ; B5 = 20 ;
% Generating is repeated until achieving
4.7% of area
while
((sum(sum(not(not(picRes))))*100)/(m*n)<4.7
)
   for(i=1:m)
    for(j=1:n)
        if(picRes(i,j)>0)
            picRes(i,j)= picRes(i,j) + 1;
            x1= i; y1= j;
            % Calculating the distance
            % of agent to 5 critical points
            d1=distance(x1,y1,A1,B1);
            d2=distance(x1,y1,A2,B2);
            d3=distance(x1,y1,A3,B3);
            d4=distance(x1,y1,A4,B4);
            d5=distance(x1,y1,A5,B5);
            % Calculating offset of new
point
            x2= (A1/(d1^2))+ (A2/(d2^3))+
(A3/(d3^3))+ (A4/(d4^2))+ (A5/(d5^2))+x1;
            y2= (B1/(d1^2))+ (B2/(d2^3))+
(B3/(d3^3))+ (B4/(d4^2))+ (B5/(d5^2))+y1;
            if(int16(x2)<62 && int16(y2)<
44 )
                % New cell must be in range
                gTemp(int16(x2),int16(y2))=
1;
            end
        end
    end
   end
   % Adding new generation after one
navigation
   % to premier matrix
   picRes=picRes+ (gTemp);
   gTemp = zeros(picSize);
end
 %% PART 3 : OUTPUT
 h=4; % Height of fixed columns
 % Calculating solar elevation angle for
local time & place
 Delta= 23.45*sind(( (daysdif('3/21/2011',
datestr(today))) *360)/365) ;
 Fi = 29.6 ; % Shiraz latitude : 29.6 N
```

```
 W= 15*(12-(hour(now))) ;

 Alfa =
asind(sind(Fi)*sind(Delta)+cosd(Fi)*cosd(De
lta)*cosd(W)) ;

 % Calculating solar azimuth angle

 Gama = asind((cosd(Delta)* sind
(W))/cosd(Alfa));

 % 'dis' is the offset distance of each
agent to

 % column which is used to prepare shadow

 dis = h/ tand(Alfa);

 % Calculating offset of shield

 dx= dis* sind(Gama);

 dy= dis* cosd(Gama);

 shadowMap = zeros(picSize);

 % Defining the situation of coulumns and
bolbs

 for(i=1:m)

    for(j=1:n)

        if(picRes(i,j)>0)

            if(int16(i+dx)<62 &&
int16(j+dy)< 44 )


shadowMap(int16(i+dx),int16(j+dy))= 1;

            end

        end

    end

 end

  %% PART 4 : SELF-CORRECTION

   % Repeate after 10 min
```

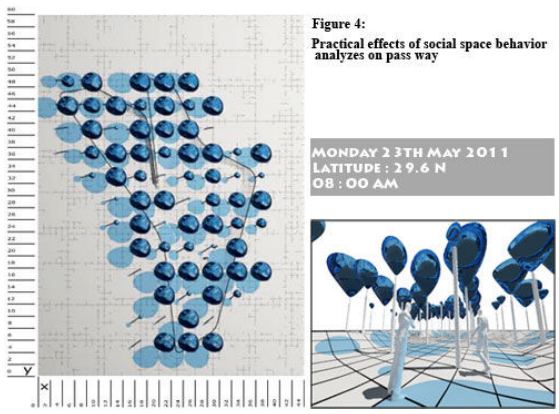
Table 1- Source Codes in "Mathlab"

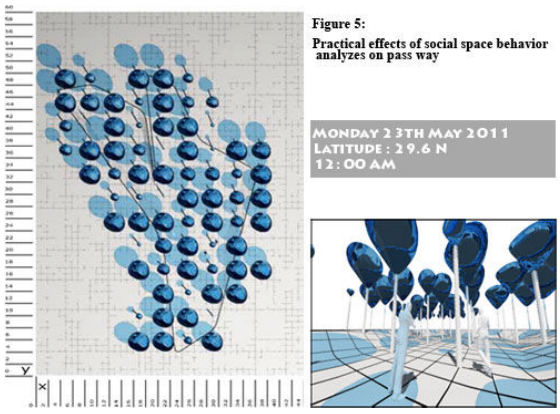
3.3 Output

A 4*4 meter grid with columns at the same height is designed on the pass way which is combined by the changeable bulbs at the end. A variable shadow is provided by the change in the capacity of the balloons. The output will be the accurate determination of full and empty balloons. In order to know that which balloon should be full to make shadow on the one square meter of the pass way (or the corresponding cell of the matrix) there is a need to solar elevation and azimuth angel. These angles are calculated for that moment and for the geographic latitude of 29.6 N and then by the application of Thales theorem position of canopies are defined and place of columns in the neighborhood of the target column is defined. At last, according to fresh data, system will full or empty the balloons with gas.

3.4 Self- Modification

In this phase, system uses self- modification to stay alive and to adapt itself to new changes of the environment. Changes include the movement of sun or even changes in the behavior of the people passing through the pass way. System will perform its modification by the repetition of the operations and renewed process of fresh data. Predetermined time for the self-modification is about 10 minutes which can be also a variable, based on the social space behavior.





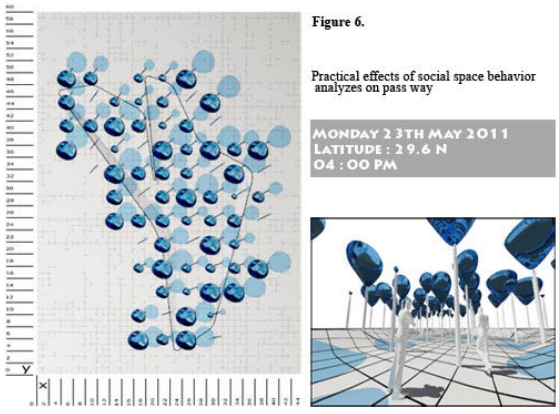


Figure 4,5,6- Urban Laboratory

4 Conclusion

In this paper authors tried to bond a realistic spatial space to theorical basis of a theory by simplifying chaotic social space behavior related variables. Methods use to monitor, analyze, generate data and geometrical functions and program algorithm are considered as variable parameters as well. So the whole issue is transforming to a meta-theory by its capability of accommodation, and this theory could be used for various situations.

5 References

[1] gleick,james,"chaos,making a new science",1987,penguin group

[2] sportt,j.c.,"strange attractores creating patterns in chaos",1993,M & T

[3] g.ivanevic,Vladimir,t.ivanevic,tijana,"high dimensional chaotic and attractor systems:a comprehensive introduction",springer

[4] reichl,l.e,"transitions to chaos",1992,springer

[5] scott,barton,"chaos.self organization and psychology"

[6] stewart,i.,"the mathematic of chaos",oxford:Blackwell

[7] f.drucker,peter,"the new productivity challenge",1991

[8] baker,g.l.,goollub,j.p.,"chaotic dynamic:an introduction"

[9] faghih,nezameddin,montazeri,m.m.,"production line balancing by genetic algorithm"2010,navid shiraz

[10] faghih,nezameddin,"dynamic systems:principles and identification",1994,semat

[11] jabbari jahromi,ali,alighanbari,azadeh,"sustainability of functional changes in real-time analyses of the socio-spatial structures",IKE2008,CSREA press

On Some Experience of Integrating Social and Semantic Web Functionality in CMS

Gurpreet Dhillon and Darina Dicheva
Computer Science Department, Winston Salem State University
Winston Salem, NC, USA

Abstract *- Recent advancements in the Web development hold promises for improving the discoverability, searchability, and shareability of resources in the digital libraries and repositories. The Semantic Web offers technologies for conceptual structuring, annotating, and indexing of resources that allow for efficient semantic retrieval while the Social Web offers means for sharing, distributing, and recommending resources. In this paper we present the results of our attempt to combine the benefits of these two technologies in a single digital repository framework by integrating Drupal - the most popular and powerful Content Management System with social features, and Fedora Commons - a Digital Repository Architecture enabling the use of Semantic Web technologies.*

Keywords: Digital Libraries, Semantic Web, Social Web, Drupal, Fedora Commons

1 Introduction

Although the number of digital libraries and repositories and the resources in them is steadily increasing, they still don't get the use they deserve. Among the main reasons for the slow uptake of the digital repositories is the challenge of keeping up with their growing number and the required efforts in searching and sharing content of interest stored in them. The recent advancements in the Web development hold promises in that direction. From one side, the Semantic Web offers technologies for conceptual structuring, annotating, and indexing of digital resources that allow for efficient semantic retrieval. From another side, the increasingly popular Social Web offers means for sharing, distributing, popularizing, promoting and recommending online resources. Therefore, a combination of both technologies holds the potential to improve the discoverability, searchability, and shareability of digital repository resources. This idea motivated us to investigate the possibilities for combining the benefits of these two technologies by employing them in the design of a framework of a Content Management System (CMS) that can be used for creating digital repositories.

We explored various approaches and software tools to implement such a framework. The most powerful and attractive solution that we found was to combine the popular content management system Drupal [1] with a Fedora Commons repository [2], which would provide our application with a Drupal-based front-end, enabling an intuitive interface and rich functionalities for supporting individual users and communities, and a Fedora-based backend that enables the use of Semantic Web technologies. The only existing software claiming to integrate Drupal and Fedora that we found was Islandora [3], so we settled on using it. To test our framework we created a site for bookmarking educational resources.

Our exploration ended up with unexpected unfavorable results. In most computer science forums, what gets published is typically success stories or positive results. Rarely do we hear about negative results or failures in pre-development research. In experimental work though, negative results are as valuable as positive, since they can help eliminate trials shown to have little or no value. Given that negative results can be useful, it seems natural that they should be also discussed and published.

The paper is organized as follows. In Section 2 we present Fedora Commons. Section 3 describes the content management system Drupal and its social features. Section 4 discusses the proposed approach to integrating Social and Semantic Web features in a content management system. In Section 5 we present our work on the implementation of a prototype of a Drupal-Fedora based Bookmark Repository, and Section 6 contains a discussion and conclusion.

2 Semantic Web and Digital Repositories: Fedora Commons

The goal of the Semantic Web [4] is to enable computers to get better in understanding and processing the data on the web and in supporting automated reasoning. This is to be achieved by enriching available information with machine-processable semantics. The aim with this effort is to bring progressively more meaning to the information published on the web. In such knowledge-based web "Automated services will improve in their capacity to assist humans in achieving their goals by "understanding" more of the content on the web, and thus providing more accurate filtering, categorization, and search of information resources." [5] The later is very important in the context of digital libraries and led to coining the term Semantic Digital Libraries [6]. A key enabler for the Semantic Web is on-line ontological support for data, information and knowledge exchange. The Semantic Web technologies include standards,

markup languages (e.g. RDF, RDFS, OWL), and related processing tools.

To our knowledge, Fedora Commons is presently the only general digital repository system, which uses Semantic Web technologies to represent and utilize the relationships between content items in its repository. This is done by maintaining and querying an RDF [7] triple store.

Fedora (Flexible Extensible Digital Object Repository Architecture) is architecture for storing, managing, and accessing digital content in the form of digital objects [8]. The Fedora Repository is very flexible; it is capable of serving as a digital content repository for a wide variety of uses. Among these are digital asset management, institutional repositories, digital archives, content management systems, scholarly publishing enterprises, and digital libraries. A Fedora Repository is able to store any sort of digital content items, such as documents, videos, data sets, computer files, and images. In addition, it allows storing of a variety of information about the content of the items (metadata).

2.1 Fedora's Digital Objects and Object-to-Object Relationships

Fedora Commons was designed to benefit from the Semantic Web technologies, such as ontologies and the W3C standards RDF, RDF Schema, OWL, and the Semantic Web Query language SPARQL. Fedora digital objects are stored in an RDF triple store. A triple store contains RDF triples and provides an efficient way for semantic querying of data and data serialization.

Fedora's ability to store data in triple stores also opens a new gateway towards using logic for reasoning. Reasoning can increase the background knowledge in the system by making inferences based on the use of classes, subclasses, properties, sub-properties, and set operations such as union and intersection in RDFS. Triple stores also support OWL, thus allowing the use of value restrictions, cardinality, transitivity, equivalence, and logical operations. Background knowledge is one of the ultimate goals of the Semantic Web and Fedora is one of the best digital repository frameworks to use for this purpose.

A Fedora repository is similar to a web content management system in its role of storing and providing access to digital content, but with a greater focus on preservation and flexibility of the content model. A Fedora repository contains Fedora objects. Unlike the traditional CMS hierarchical content models, Fedora objects are structured as a graph of content nodes.

A Fedora object contains several *DataStreams*. A data stream is a component of a digital object that represents a data source; thus the object's data streams contain the actual data. This data can be textual, audio-visual or metadata about the object (for example, stored in an XML file). Every object has a reserved Dublin Core data stream that is created by the Fedora repository service automatically if one is not provided. In addition to the reserved data streams, users can add as many data streams to an object as needed.

Fedora digital objects can be related to other Fedora objects in many ways. In the Content Model Architecture (CMA) object serialization these relations are asserted as RDF statements in the Fedora objects' RELS-EXT data stream (see Fig. 1).

Subject	Predicate	Object
info:fedora/islandora:36	rel:isMemberOf	info:fedora/islandora:22
info:fedora/islandora:36	fedora-model:hasModel	info:fedora/bkmrk:cm

Figure 1. RDF triples with Subject, Predicate and Object

3 Social Web and CMS: Drupal

Drupal is one of the most popular content management systems nowadays. It allows users to easily organize, publish, and manage a wide variety of content on a website. It is a free and open-source content management system and content management framework distributed under the GNU General Public License. Drupal has built a reputation of innovation, stability and flexibility. Drupal boasts thousands of modules for almost any type of digital presence, with a focus on personalization, community building and social tools. The standard release of Drupal, known as Drupal Core, contains basic features common to content management systems. In addition to content management, these include page layout customization, menu management, user account registration and maintenance, and system administration.

3.1 Drupal's Social Features

By design, Drupal includes a number of the popular social web sites' features:

- Commenting – Authorized users can comment on any Drupal document, as well as reply to a comment or edit a comment (own or somebody else's depending on their permission).
- Tagging – Authorized users can tag any Drupal document, such as page, story, book, or blog entry. In addition, they can classify them to predefined categories.
- RSS – Through RSS users can get latest updates about changes in Drupal documents, such as page, story, book, or blog entry.
- Blogs – Drupal supports the creation of blogs. Blogs present a way of publishing and updating news on the web. Typically a blog post/entry contains the title and the author of the topic, a time stamp, a description of the topic, and optional elements, such as comments and rating. Blog entries can be displayed in an ascending or descending order on the page. In Drupal, users with an appropriate authorization can create and maintain blogs that are private or public to the site members; authorized users can then create and comment on blog posts.

In addition, many Drupal modules providing other Social Web features have been created and contributed by the

Drupal community. These include among others: rating, voting, the Facebook 'Like' button, a 'Share' button that allows users to share a webpage content by email and to sites like Facebook, Twitter, Diggs, etc.

Users can also create wikis in Drupal. Although there is not a Drupal content type explicitly corresponding to a wiki, one can create a wiki by creating a blog and using the functionality provided by the Drupal's 'Organic Groups' module. The idea is to give the permission for creating blog posts to a community (an organic group) as opposed to a single user (the blog owner).

4 Integration of Social and Semantic Web Features in CMS

There are different ways for integrating Social and Semantic Web functionalities in CMS. Apparently, the most efficient one is to re-use an already existing CMS framework. The problem is that currently such a framework could at most provide either Semantic Web or Web 2.0 functionalities. Thus, if we consider our selected representative CMS, Fedora Commons and Drupal, there are three clear ways for such integration:

- Drupal-based CMS: This approach requires developing a Drupal module that supports semantic annotating and querying of content.
- Fedora-based CMS: This approach requires extending the Fedora Commons' framework and Fedora's data model with social features.
- Drupal-Fedora based CMS: This approach requires integrating a Fedora repository within a Drupal installation.

We chose the last option since it appeared the most powerful and implementation-wise efficient. In order to implement the integration, we had to develop 'connecting' software either from scratch or by finding open source software to use as a basis. Our extensive search resulted in finding only one tool – Islandora – an open source software developed by the University of Prince Edward Island's Robertson Library [3]. According to the developers "Islandora is aimed at combining the Drupal and Fedora open software applications to create a robust digital asset management system that can be fitted to meet the short and long term requirements of digital repositories." Thus Islandora seemed to be appropriate for achieving our goal of integrating Social and Semantic Web functionalities in CMS.

The integration of Drupal and Fedora through the use of Islandora includes installing Drupal, installing Fedora Commons, installing Islandora, and configuring Islandora. Islandora uses third party software products and thus depends on many other installations. The installation process was quite cumbersome and in order to share our experience and help other researchers and practitioners avoid installation problems, we created detailed installation instructions (posted at http://iiscs.wssu.edu/ drupal/ ?q=node/388).

The Drupal Islandora module allows Drupal users to view and manage digital objects stored in the Fedora repository. It comes with features like bulk ingest, collection management utilities and streamlined solution packs. Authorized users can configure Islandora's various properties from the Drupal's Site Configuration menu. Registered and authorized users can add, delete, search and manage digital objects. Anonymous users are only able to search digital content. Islandora Solution Packs are small Drupal modules offering custom content model objects, workflows, and ingest forms (the forms a user sees when ingesting a new object into the repository). The solution packs present a starting point for users by providing means for using particular types of data (such as books or audio files), based on typical experiences of working with these types of data. Drupal users can install different solution packs into a repository by going to Islandora's Configuration Pane and selecting the specific solution pack.

5 Prototype Implementation of a Drupal - Fedora Based Bookmark Repository

Our goal in implementing the Bookmark Repository prototype was to experiment to what an extent a repository, created in the framework that integrates Drupal with Fedora Commons, can benefit from the underlying Social and Semantic Web functionalities.

The application was required to offer the standard functionalities for a social bookmarking website: for registered users, the ability to maintain (create, edit, and delete) their bookmarks, search and navigate other users' bookmarks, post comments, tag bookmarks, rate bookmarks, and receive RSS feeds. The registered users should be also able to create groups for sharing bookmarks with the other group members.

From the semantic perspective, we envisaged building an ontology, which would be used as a controlled vocabulary for tags as well as for semantic browsing and search of the content.

The implementation included the following steps:

1. Installing Drupal.
2. Installing a Fedora repository.
3. Installing Islandora.
4. Configuring Drupal user roles and Islandora templates for communication and assigning role privileges.
5. Creating a Fedora content model for the Bookmark collection.
6. Creating a form for entering data in the Bookmark collection.
7. Creating the Bookmark collection and entering data.

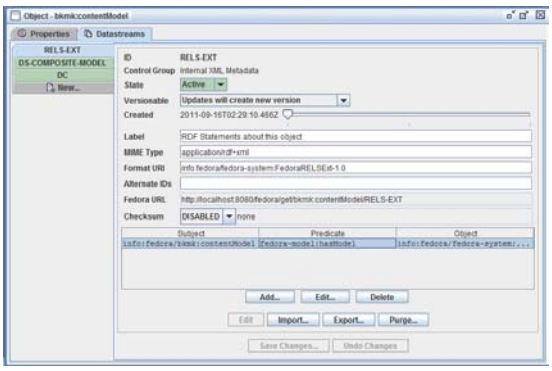


Figure 2. Fedora's Bookmark content model

Figures 2, 3 and 4 show screenshots of the Fedora and Drupal user interfaces related to some actions in completing steps (5), (6), and (7), correspondingly.

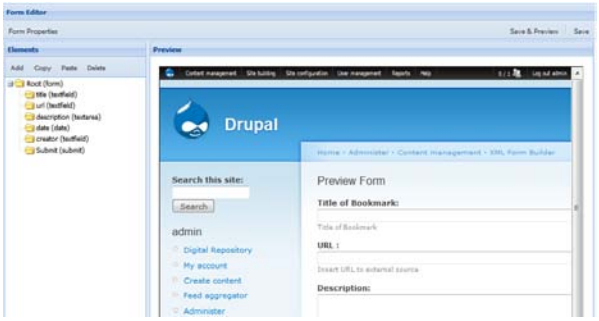


Figure 3. Defining a bookmark form by using the Form Builder

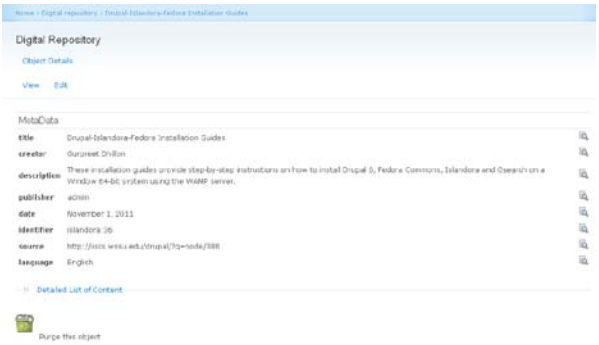


Figure 4. Metadata of an ingested bookmark object

A major problem in installing and using Islandora was that the project was (and still is) under active development, which implied frequently changing code and lack of stable and detailed enough documentation. Many features in the provided documentation did not work or were under development.

6 Discussion and Conclusion

Fedora Commons is a framework that supports a semantically rich repository organized as an RDF triple store, which can be manipulated or queried to help providing meaning to the data. On the other hand, Drupal is a content management system, which provides an easy way of creating websites and a number of Social Web features. The goal of this work was to investigate the possibility of integrating the functionalities of both platforms targeting a framework that benefits from both the Social and Semantic Web technologies. As a middle layer for combing Fedora and Drupal, we used Islandora, the only existing software claiming to integrate them. Using the three open source software systems, we created a prototype web application using Drupal as a front-end and a Fedora repository as a back-end. In this framework we implemented the Bookmark Repository as a test bed for experimenting to what an extent such an application can benefit from the underlying social and Semantic Web functionalities of Drupal and Fedora.

Our experience shows that although using Islandora saves the need for implementing a middle layer for connecting Drupal to Fedora, its installation is very difficult and requires extensive knowledge of a number of technologies, as well as expertise with various software products. The resulting application can benefit from the rich Fedora functionalities, since the information is stored in a Fedora repository however extensive programming and a very detailed knowledge of Fedora are needed for creating the different data streams that can support the processing and rendering of information from the repository.

Our biggest disappointment was that we were not able to use, as initially hoped, neither contributed Drupal modules that support social/community features, such as rating, voting, sharing, etc., nor even the basic Drupal features, such as tagging and RSS.

Since the access to the information stored in the Fedora repository is only through the Islandora interface and this interface is completely separate from the main Drupal interface, there is no way to apply Drupal core functions to objects stored in the Fedora repository. We could not use any contributed Drupal module either, since Drupal modules are designed and configured to process only data stored in the Drupal database and apparently Islandora does not provide the necessary bridge to the Fedora database. It seems that the only way for Islandora to support some of the Drupal's social features is to re-implement them.

There was no way to realize this lack of proper integration between Drupal and Fedora before the completion of the complicated installations of Fedora and Islandora, the Islandora's configuration, and the actual creation of a digital repository (in our case the Bookmark repository). Taking into account the importance of the problem and the popularity of the involved software systems (thus the likelihood of repeating the study by others) and the substantial time needed to get to this conclusion, we consider that sharing this result although negative, can be valuable to many researchers and practitioners.

In conclusion, our research suggests that a better approach to enhancing a Drupal-based content management system with Semantic Web features would be to keep the data in the Drupal database and create a native Drupal module,

which extends the system with the desired semantic features. We are planning to pursue this direction in our future work.

7 Acknowledgments

This material is based upon work supported partly by the NSF Grant No. DUE-1044224 "NSDL: Social Bookmarking for Digital Libraries: Improving Resource Sharing and Discoverability".

8 References

[1] Drupal – Open Source CMS. http:// http://drupal.org/.

[2] Sandra Payette and Carl Lagoze. Flexible and Extensible Digital Object and Repository Architecture (FEDORA), *Proceedings of ESDL 98*, 1998.
Available at
http://web.archive.org/web/20000915230310/www.cs.cornell.edu/payette/papers/ECDL98/FEDORA.html.

[3] Paul Pond, Peter MacDonald. Islandora Guide, 2011, Available at
https://wiki.duraspace.org/display/ ISLANDORA/Islandora+Guide

[4] Tim Berners-Lee, James Hendler and Ora Lassila. The Semantic Web, *Scientific American*, May 17, 2001.

[5] Ding, Y., Fensel, D., Klein, M. and Omelayenko, B. The Semantic Web: Yet Another Hip? *Data and Knowledge Engineering*, 41(3): 205-227, 2002.

[6] Sebastian Kruk. Semantic Digital Libraries, Tutorial given at *International Conference for Digital Libraries ICDL'2009*, 2009, Available at
http://www.slideshare.net/knowledgehives/tutorial-on-semantic-digital-libraries-at-icsd09

[7] RDF/XML Syntax Specification, W3C Recommendation 10 February 2004. Available at
http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/#section-Syntax-node-property-elements

[8] Ron Daniel Jr., Carl Lagoze, Sandra D. Payette. A Metadata Architecture for Digital Libraries, *Internet Archive*, 1998,
Available at
http://web.archive.org/web/20000915230436/www.cs.cornell.edu/lagoze/papers/ADL98/dar-adl.html.

A Hardware-assisted Instruction Security Monitoring Design in Embedded System

Zichen Zhou, Bin Xu, Qining Lu, Bo Yin, Renhao Fan, Tao Liu, and Xiang Wang
School of Electronic and Information Engineering, Beihang University, Beijing, 100191, China

Abstract: *This paper presents a series of novel architectural-enhanced security solutions. In the cross-compilation link stage, the automated compiler extracts the intrusion model for instruction code and static data, meanwhile secure tags of each main memory segment are added at the compile time automatically. At runtime, the designed hardware observes its dynamic execution trace and checks whether the trace conforms to the permissible behavior and trigger appropriate response mechanisms. The proposed methods don't change the compiler or the existing instruction set and imposes no special restriction to the software developer. The hardware structure of protection design is implemented on an actual OR1200-FPGA platform. The experimental analysis shows that the proposed techniques can prevent a wide range of common software and physical attacks with minimal resource cost and low performance consumption.*

Keywords: Secure tag; instruction security; architectural-enhanced; embedded system

1. Introduction

With the development of technological innovations in previous decades, embedded systems penetrate deeper into our lives and are often required to deal with sensitive information or perform critical functions such as communication network and automotive control [1].

However, most embedded software expose a number of vulnerabilities [2]. Additionally, with the use of networks through wireless or the internet, many embedded systems expose themselves to potentially vulnerable programs and becoming easy targets for software-based security attacks. One of the most common forms is buffer overflow attack [3].

Furthermore, the ubiquitous use of embedded systems makes it easier for an attacker to gain physical access. By using advanced electronic equipment the sophisticated attacker can launch physical attacks to tamper or modify the instruction code, interrupt the communication of the processor and control the execution in his desired direction [4].

Most of the existing approaches tackle the security problems at the software level, but they can't avoid vulnerabilities and often induce high overhead in performance [5].

This paper presents a novel hardware solution to enhance the application execution security which can be summarized as follows. The off-line software extracts the control flow integrity monitoring model at compile time. At runtime, the designed hardware observes its dynamic execution trace and ensures the application does not deviate from its intended behavior. When a mismatch is detected, the hardware will trigger the response mechanism [6].

The rest of the paper is organized as below. Section2 presents the instruction model architecture in detail. A security analysis of our mechanism is provided in Section3. Section4 shows the experimental result and security analysis. At last, conclusion was drawn by Section 5.

2. Hardware-assisted Secure Monitoring Architecture

This section provides a hardware-assisted instructions monitoring for intrusion detection to prevent the implementation of wrong control flow.

Hardware-assisted instructions monitoring for intrusion detection: The paper proposes a novel, hardware-assisted, hierarchical framework for monitoring the control flow of embedded programs at a fine granularity, and a methodology for designing and configuring the hardware monitors. The framework catches all attacks on programs that operate by altering its control flow.

2.1 Architecture Overview

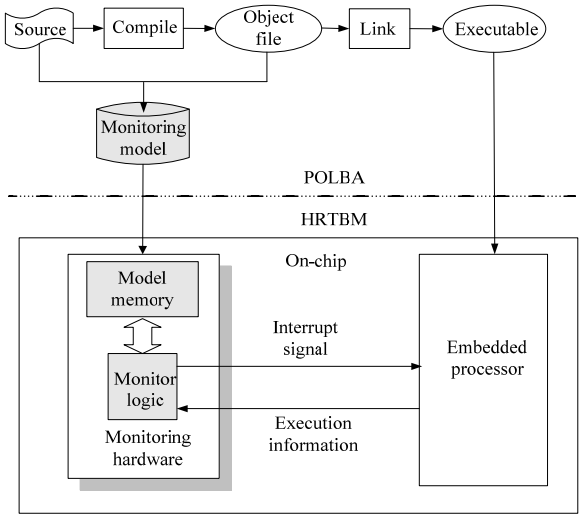


Figure 1 instruction models architecture

The instructions models architecture we propose is shown in Figure 1. There are two parts in this architecture: program off-line behavior analysis (POLBA) and hardware real-time behavior monitoring （HRTBM）.

In POLBA, the source program is cross compiled and linked to generate the executable binary code. Meanwhile, the program is analyzed to extract a control flow integrity monitoring model. The executable binary code is stored in the CPU main memory while the monitor model information is stored in the on-chip ROM memory. In HRTBM, the runtime execution information is sent to the monitoring hardware. The hardware logic compares the runtime execution information stream with the monitoring model and checks whether the execution trace is permitted. Once the runtime execution behavior is not permitted, a control signal will trigger the response mechanism (e.g., terminating or recovering the program). In this system, the monitoring hardware is implemented on-chip. So it can not be compromised by many malicious software and physical attacks.

2.2 Monitor Model

A good monitoring model needs to have the capacities as follows: a) easily extracted through automatic program analysis for a wide range of programs. b) Accurately describe the program intended control flow integrity behavior. c) Promptly detect the unintended behavior. d) Necessarily meet lightweight demand for resource-limited embedded system.

Considering the above requirements, the basic block level is chosen to monitor the program execution. We define flow control instructions, such as branch and jump, indicate the end of a basic block, and the next instruction to be executed is the beginning of another basic block. Our monitoring model depends on the control flow graph. It contains three sets: F, B&I.

F: Function calling information of program.

$$f_i \in F,\ f_i = \{address\} \tag{1}$$

fi is the ith function absolute entry address.

B: Basic block jump information of program.

$$b_i \in B,$$

$$B_j = \begin{Bmatrix} index[j], addr\_b[j], \\ type\_b[j], TARGET\_b[j] \end{Bmatrix} \tag{2}$$

bj is the jth basic block jump information. index_f[j] is the jth basic block corresponding function index in F; addr_bn[j] is the relative address between the entry address of the jth basic block and the entry address of the function corresponding to the jth basic block; type_b[j] is the jumping type of the jth basic block which is shown in TABLE 1. TARGET_b[j] is the possible target jump address sets of the jth basic block. Table 1 shows the next possible path of control flow. After this section, the application basic blocks are extracted. So the control flow is known.

I: The application code integrity information

$$I_i \in \{AES / LFSR\} \tag{3}$$

Ij is the jth basic block integrity information of application code. Its value is calculated by AES/LFSR algorithms.

Based on the F, B and I sets, the instructions model is generated.

Table 1 Path of control flow

X	Type_b	Addr_t
0	Intra-function unconditional jump	$\forall$ addr_t $\in$ TARGET,addr_t=addr_f[i]+addr_bn[t];t $\in$ TARGET_b[j]
1	Intra-function conditional jump	$\forall$ addr_t $\in$ TARGET, addr_t=addr_f[i]+addr_bn[t]; or addr_t=addr_f[index_f[j+1]]]+addr_bn[j+1]; $\in$ TARGET_b[j]
2	Inter-function unconditional jump	$\forall$ addr_t $\in$ TARGET, addr_t=addr_f[index_f[t]]]+addr_bn[t]; t $\in$ TARGET_b[j]
3	Inter-function conditional jump	$\forall$ addr_t $\in$ TARGET, addr_t=addr_f[index_f[t]]]+addr_bn[t]; or addr_t=addr_f[index_f[j+1]]]+addr_bn[j+1]; t $\in$ TARGET_b[i]
4	Multi branches jump	$\forall$ addr_t $\in$ TARGET, having an traverse search of TARGET
5	Function call return	addr_t obtained directly through a stack mapped unit (SCS)

The statically extracted monitoring model is loaded into an internal F, B & I memory table and the monitoring hardware is embedded into the pipeline stages to perform the runtime checking as in Figure 2 shows.

During program execution, the hardware buffers current basic block's beginning Program Counter (PC) signal at IF stage. Depending on this, the monitoring logic fetches the corresponding basic block's monitoring model (F, B and I value). Then, the monitoring hardware logic begins to calculate the next possible control flow address information. When the next control flow instruction is encountered, the hardware can compare the next basic block's real time PC with the calculated one to validate the run time control flow behavior. At the same stored in the monitoring model to validate the integrity. The basic blocks are imposed a limit length who excess must be separated into sub-blocks and an XOR of the hash values is checked.

The compare unit detects deviation of program execution at run-time. When the checking is not accomplished, the processor must be freeze in order to allow the compare unit to catch up. When founding a mismatch, the monitor logic will assert the control signals to notify the Operating System to trigger response mechanism.

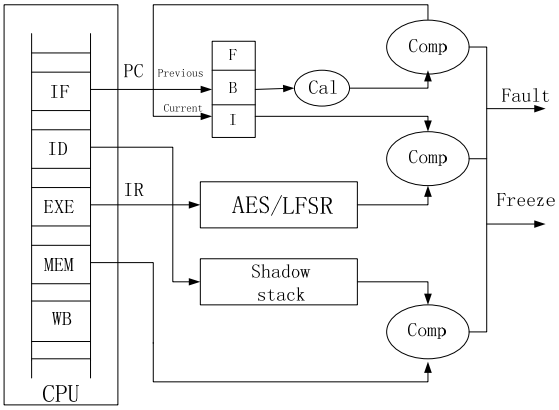


Figure 2 The architecture of the proposed hardware monitor

Because the monitoring model concerns a finite, static CFG, it can not be used to calculate the return address of the calling function. A shadow call stack (SCS) is employed to provide correct, uncorrupted return addresses to settle this problem.

In ID stage, the hardware SCS decodes the instruction. When founding a push instruction that store the return address onto the stack, the SCS pushes the current return address block information onto the top of its corresponding last in first out (LIFO) memory. When founding that there is a pop instruction that load the return address the SCS pops the current return address block information from the top of LIFO memory. In MEM stage, the return address from the stack memory is popped out. Then the compare unit determines whether it equals to the return address popped from the SCS and trigger the response mechanism.

2.3 Secure Tag Validation

All application processes are loaded into three major memory areas: the stack segment, the data segment, and the code segment as in Figure 3 shows.

The stack segment contains stack and heap. The stack stores the local variables and procedure calls. Heap, similar to stack, is a region of virtual memory used by applications. However, unlike stack, private heap space can be created and freed by programmers. The data segment stores static variables and dynamic variables, and it contains Data, BSS, Rodata[7]. Data is used to the initialed data, Block Started by Symbol (BSS) is used to store the un-initialized data, and Rodata is used to store the read only data. The code segment contains Code and Vector. Code stores the program instructions and Vector stores the exception handlers. Each segment has different attributes of writable readable and executable as shown in Figure 3.

Many buffer overflow attacks inject malicious code in the data segment. For avoiding this kind of attack, we can tag the data segment un-executable. When the CPU executes instructions in the data segments, the attack will be detected. Similarly, the stack segment is private to each application, and no other application can access those areas. Few applications are designed to execute in the stack. Only in a little conditions，the OS permits executing the code in the stack segment. In this paper, we define the stack is un-executable and only the code and vector segment is executable. The code portion on the other hand is a read-only segment. However, if an attempt is made to write to this area, a segment violation occurs. So we define only the Data, BSS, stack and heap segments are writable. If the CPU loads the application data in the code segment, it will also lead to a violation. So, we define only the Data, BSS and Rodata segments are readable.

Two bits of memory width are used to describe the attributes of the above segments. We define 10 presents the executable, un-writable and un-readable, 00 presents the un-executable, un-writable, readable, and 01 presents the un-executable, writable and un-readable attributes as in figure shows.

10	Vector
10	Code
01	Data
01	BSS
00	Rodata
01	Stack
01	Heap

10	Executable	Un-writable	Un-readable
00	Un-executable	Un-writable	Readable
01	Un-executable	Writable	Readable

Figure 3 The attributes of each segment

At compile time, the compiler generates the secure executable binary code. And the secure tag will be extracted according to our definition. In our design, the binary code is stored in the main memory and the secure tag is loaded in an on-chip memory which can't be accessed by the attacker. So it's can't be tampered by software or physical attacks as in Figure 4 shows.

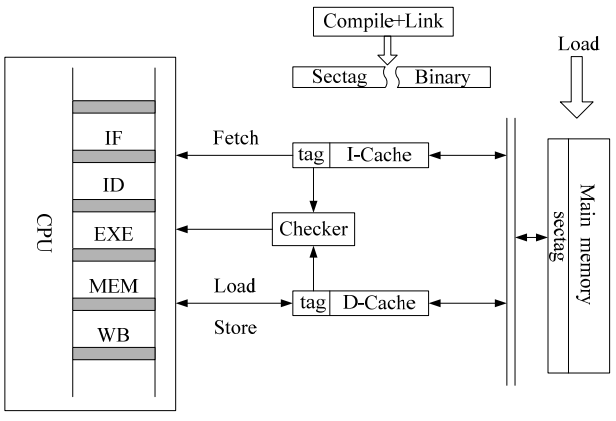


Figure 4 The secure tag architecture

The second is on the readable attribute. The data is sent to the CPU at the MEM stage when executing the Load instruction. When a D-cache miss occurs, the application data is sent to the D-cache from the main memory. The checker receives the corresponding secure tag and check if it's from the readable region.

The third is on the writable attribute. The application data can be write back to the main memory when execute the Store instruction at MEM stage. Depending on the writing target address, the checker can determine whether the memory region is writable.

When finding the execution doesn't conform to our designed attributes, the checker can send the interrupt signal to the CPU.

2.4 Secure Memory Access Validation

Not all the data can be fixed off-chip, and there is variable data generated during the execution and change from time to time. That is a drawback of static analytical intrusion detecting model. In order to solve the problem, we present a dynamic protection for those data which access the insecure memory and possible suffer from the physical attack [8].

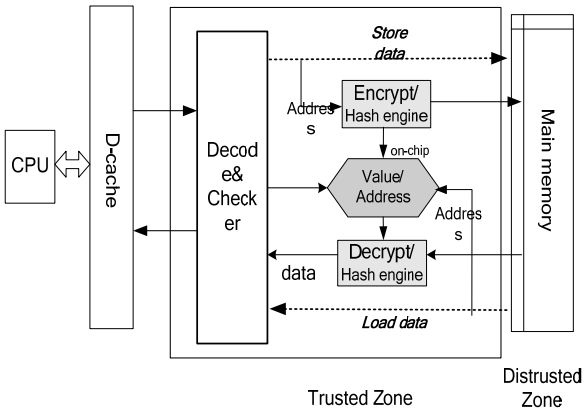


Figure 5 Memory access detector architecture

At runtime, when the detector firstly decodes IR instructions and finds out a memory access instruction, when it makes sure the instructions will access main memory which we has defined it as distrusted zone, and checker send an enable signal to make the monitor work. What's more, it does a further judgment that the instructions belongs to a read or write memory access instruction. If it means writing, the monitor receives memory contents before the CPU and encrypts it depending on the cache block. These confidentiality data is stored at our on-chip memory. When executing a read instruction, if a data cache miss occurs, the CPU issues an enable signal and then the memory dumps the contents on the bus. The hardware decrypts the data before sending into the D-Cache. The hardware is aware of the data address requested, so it can fetch the previous stored data.

If there is a mismatch, it depicts that the application data is tampered on off-chip memory. As a result of potential attacks, the checker will launch a freeze signal to cease the current execution and make the processor jump to designated address where protective codes load there.

Many attacks are initiated by tampering the application data rather than altering process flow or tampering application code. In order to detect these physical attacks on unsafe off-chip main memory data, we design the data confidentiality validation hardware as in Figure 5 shows. The information of static data can be derived at compile. However the run-time data can be fetched only while the process is active. As Figure 5 shows the hardware monitors all memory communication between the CPU and main memory. At runtime, when the detector finds out a memory access instruction, the hardware decodes the instruction and make sure it will read or write the memory. If it means writing the monitor receives memory contents before the CPU and encrypt/Hash it depending on the cache block. These values are stored at our on-chip memory. When executing a read instruction, if a data cache miss occurs, the CPU issues an enable signal and then the memory dumps the contents on the bus. The hardware decrypt/Hash the data before sending into the D-Cache. The hardware is aware of the data address requested, so it can fetch the previous stored data. If there is a mismatch, it depicts that the application data is tampered on off-chip memory.

3. Security analyze

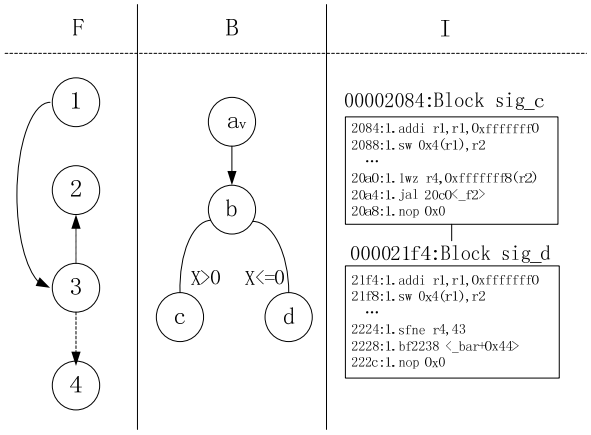


Figure 6 The monitoring granularities of the security analyze

The designed architecture can detect any attacks that tamper the application control flow integrity no matter by software or physical attacks. It can protect the application execution from three granularities. The first one is on the function level. As in figure 6F shows, the normal function call flow is①→ ③→ ②, with a malicious attack that change the flow to ①→ ③→④, and the④may be a malicious code. The second one is the basic block level. As in figure 6B shows, there are 5

basic blocks in function 1. Our mechanism can protect the basic block jumping different from the statically extracted monitoring model. Now consider the branch instructions in b. For example, in if (x>0) b→c, if (x<=0) b→d. c and d are both considered to be the effective flow transferring from b. But if the attacker tampers the variable x, he can control the execution flow easily. The third one is based on hash signatures for all the basic blocks. As in figure 6I shows, the 20c0's corresponding signature is sig_c, 20e8's corresponding signature is sig_d. Suppose the processor requests (address for) block c after block b, the attacker can tamper, inject code in block_b or substitute it with block d. The monitoring hardware pick up sig_c by reading the entry address and compute the signature of the requested one. Since the signatures are obviously different, the hardware can detect the attack. For replay attacks, our mechanism makes sure that the signature is the latest one.

4. Experimental Results

The embedded processor adopted in this paper is OR1200 [9] which is a 32-bit scalar RISC with Harvard micro architecture, 5 stage integer pipeline, virtual memory support (MMU) and basic DSP capabilities. The system on programmable chip (SOPC) platform is built on Altera FPGA. The software development tool for OR1200 is the popular and free GNU [9].

Table 2 Overhead and detecting speed

Set	F&B	SCS	I	
			AES	LFSR
Size(bit)	**32**	**32**	**256**	**120**
Slices	**297**	**305**	**423**	**96**
Slice Flip Flops	**253**	**264**	**274**	**160**
4 input LUTs	**569**	**583**	**797**	**76**
Clk	**⩽4**	**1**	**10**	**1**

The hardware overhead contains memory overhead for monitoring model and logic overhead for control and validation [10]. Our monitoring model memory overhead is different for diverse applications. The detecting speed of the monitoring hardware can be wagged when different algorithms are used. Once the processing speed of the monitoring hardware is lower than the processor, the processor should be frozen. In our scheme, as in Table 2 shows, the overhead of the process flow behavior detection depends on the specific executables programs and the instructions it has. We experience AES as a standard algorithm, and it will take 10 cycles for an encryption of crucial instruction. Considering the demand of run-timing, we choose a stream cipher which based on LFSR as another candidate algorithm, it will take just one cycle to encrypt every time. The main overhead should also consider the ratio of the crucial instructions to all possible instructions. The hardware can accomplish the validation within one clock in instruction fetching and data loading.

5. Conclusions

The concept of designing hardware to secure program execution in untrustworthy environments appears to have considerable potential advantage. Specialized architectural enhanced schemes protect the application from a class of software and physical attacks in embedded processors. The compiler generates the monitoring model at compile time and the hardware monitor verifies the execution trace at runtime. The detecting speed is much quicker than the software based schemes. The architectural design is implemented on an actual OR1200-FPGA platform with the associated costs in terms of extra logic and on-chip storage. Different security schemes are formulated to meet the tradeoffs between overhead and security. In conclusion, since the scheme is performance tolerant and resource overhead acceptable, this hardware-assisted secure architecture is suitable for embedded system whose resource is limited and security is demanding.

6. Acknowledgment

This research is supported by Astronautic innovation Fund and Key Fund (Grant No. CASC200902), Astronautic support Fund (Grant No. 374007), Aeronautic Joint Fund (Grant No. 2008ZC12004), the National Science Foundation of China (Grant No. 60973106), and National 863 Project of China (Grant No. 2011AA010404).

7. References

[1] Z. Shao, C. Xue, Q. Zhuge, M. Qiu, "Security Protection and Checking for Embedded System Integration against Buffer Overflow Attacks via Hardware/Software," IEEE transaction on computers, vol. 55, pp. 443–453, 2006.

[2] Arkin, Brad, Stender, Scott, McGraw, Gary. Software penetration testing. IEEE Security and Privacy, 3(1): 84-87, 2011

[3] J. P. McGregor, D. K. Karig, Z. Shi, and R. B. Lee, "A processor architecture defense against buffer overflow attacks," in Proc. Int. Conf. on Information Technology: Research and Education, pp. 243–250, 2003.

[4] G. E. Suh, D. Clarke, B. Gassend, M. van Dijk, and S. Devadas, "AEGIS: Architecture for tamper-evident and tamper resistant processing," In Proc. Int. Conf. on Supercomputing, p.160–171, 2003.

[5] A. M. Fiskiran, R. B. Lee, "Runtime execution monitoring (REM) to detect and prevent malicious code execution," Computer Design: VLSI in Computers and Processors, 2004. ICCD, pp. 452–457, 2004.

[6] X. Wang, W. Lei, L. Wang, C. Zhang, R. Guo, Z. Liu, "Hardware Monitoring to Enhance Embedded System Security," Proc of Academic Forum for 2008 China

Information Technology, Beijing: Publishing House of Electronics Industry, pp. 650–655, 2009.

[7] D. Arora, R. Srivaths, A. Raghunathan, N. K Jha, "Hardware-assisted run-time monitoring for secure program execution on embedded processors," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 14, pp. 1295–1308, 2006.

[8] C. Bu, X. Wang, C. Zhang, J. Liu, X. Wang, C. Qi, X. Gao, B. Li, "Compiler / Hardware Assisted Application Code and Data Security in Embedded Systems," 28th IEEE DASC, 25–29 October, 2009, pp. 7.E.2-1--8, and 2009 IEEE/AIAA 28th DIGITAL AVIONCS SYSTEMS CONFERENCE, Vol. 1-3, pp. 1757-1764, 2009.

[9] OpenRISC 1200 IP Core Specification. Damjan Lampret. http://www.opencores.org

[10] Aaraj N., Raghunathan A., Jha NK. A Framework for Defending Embedded Systems Against Software Attacks. ACM TRANSACTIONS ON EMBEDDED COMPUTING SYSTEMS, 2011(10):177-18

Adaptive Nonparametric Discriminant Analysis

Smarajit Bose, Amita Pal, Rita SahaRay and Jitadeepa Nayak
Bayesian & Interdisciplinary Research Unit, Indian Statistical Institute,
203 B. T. Road, Kolkata-700 108, India.

Abstract— *In Fisher's linear discriminant analysis, the assumption of equality of the dispersion matrices of different classes leads to a classification rule based on minimum Mahalanobis distance from the class centers. However without this assumption, the resulting quadratic discriminant classifier involves the Mahalanobis distance as well as a factor based on the ratio of the determinants of the dispersion matrices. It turns out that analyses with some other elliptically symmetric distributions also involve similar factors in addition to the Mahalanobis distance.*

In this paper, we propose a nonparametric technique which generalizes Fisher's discriminant analysis for a range of elliptically symmetric distributions. We present an extensive simulation study to illustrate the potential of the method. Using a variety of real life data sets we show that this generalized Fisher's Discrminant analysis is very competitive with other nonparametric methods.

Keywords: Fisher's Discriminant analysis, Mahalanobis Distance.

1. Introduction

In traditional Fisher's Discriminant Analysis, the class densities are assumed to follow multivariate normal distribution with different mean vectors for different classes. The dispersion matrices are often assumed to be the same across all classes which leads to Linear Discriminant Analysis (LDA). The resulting classification rule is equivalent to assigning an observation to the class for which the Mahalanobis distance between the observation and the class mean is minimum. We will refer to this classification rule as Minimum Mahalanobis Distance (MMD) classification rule. When the assumption of equality of the class dispersion matrices is omitted, the resulting rule is called the Quadratic Discriminant Analysis (QDA). However the resulting classification rule is not the same as MMD rule in this case as it also involves a factor based on the ratio of the determinants of the dispersion matrices. Anderson [**?**] provides an excellent introduction to Fisher's Discriminant Analysis.

Though being a popular choice for classification, QDA does not perform very well when the class densities are very different from normal distribution. In some such situations, MMD which is a purely nonparametric but simple method produces better result. In this paper we try to establish a general framework of which both QDA and MMD are special cases. Through extensive experimentation we illustrate the effectiveness of the proposed method, and also compare its performance with some established powerful nonparametric classifiers in a variety of examples. The organization of this paper is as follows:

In Section 2, we propose a new method which generalizes QDA and MMD classifiers and illustrate it with some simulated and real data in 2 class problems. Generalization of the method for multi-class (more than two classes) is presented in Section 3 which also contains the results of several experiments. In Section 4 we compare the performance of the proposed method with that of some standard multivariate classifiers. Finally concluding remarks are given in Section 5.

2. Two Class Classification:

In a two class classification problem, measurements $\boldsymbol{x} = (x_1, \ldots, x_p)$ are taken on a single individual (or object) and the individual is to be classified into one of the two classes, say π_1 and π_2 having probability density function $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$ respectively. Let there be a sample of N observations need to be classified and R_1 be the set of $\boldsymbol{x}$ values for which we classify the objects in π_1 and R_2 be the set of remaining $\boldsymbol{x}$ values for which we classify the objects in π_2. Then the decision rule is defined as,

$$R_1 : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq 1$$

$$R_2 : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < 1.$$

Considering the case of multivariate normal densities with mean vectors $\boldsymbol{\mu_1}$, $\boldsymbol{\mu_2}$ and dispersion matrices Σ_1 and Σ_2 respectively, we get the QDA rule,

$$\begin{aligned} log\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} &= \frac{1}{2}.log\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right) + \frac{1}{2}[(\boldsymbol{x}-\boldsymbol{\mu_2})'\Sigma_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu_2}) \\ &\quad -(\boldsymbol{x}-\boldsymbol{\mu_1})'\Sigma_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu_1})] \\ &= \frac{1}{2}.log\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right) + \frac{1}{2}\Delta^2, \quad \text{say}, \end{aligned} \tag{1}$$

which simplifies to:

$$\begin{aligned} R_1 &: \frac{1}{2}.log\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right) + \frac{1}{2}\Delta^2 > 0. \\ i.e. &\quad \Delta^2 > log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right). \end{aligned} \tag{2}$$

Now, if the classes are assumed to follow p-variate t distributions having n d.f., then with

$$f_i(\boldsymbol{x}) = A.|\Sigma_i|^{-\frac{1}{2}}\left[1+\frac{1}{n}(\boldsymbol{x}-\boldsymbol{\mu_i})'\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu_i})\right]^{-\frac{n+p}{2}} \quad i=1,2$$

where

$$A = \frac{\Gamma\left(\frac{n+p}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}.n^{p/2}.\pi^{p/2},$$

we get

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} = \frac{|\Sigma_2|^{1/2}}{|\Sigma_1|^{1/2}}\frac{\left[1+\frac{1}{n}(\boldsymbol{x}-\boldsymbol{\mu_2})'\Sigma_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu_2})\right]^{\frac{n+p}{2}}}{\left[1+\frac{1}{n}(\boldsymbol{x}-\boldsymbol{\mu_1})'\Sigma_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu_1})\right]^{\frac{n+p}{2}}} \quad (3)$$

Thus,

$$\begin{aligned} log\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} &= \frac{1}{2}(log|\Sigma_2| - log|\Sigma_1|) \\ &\quad +\frac{n+p}{2}[log\{1+\frac{1}{n}(\boldsymbol{x}-\boldsymbol{\mu_2})'\Sigma_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu_2})\} \\ &\quad -log\{1+\frac{1}{n}(\boldsymbol{x}-\boldsymbol{\mu_1})'\Sigma_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu_1})\}] \\ &\cong \frac{1}{2}log\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right)+\frac{n+p}{2n}[(\boldsymbol{x}-\boldsymbol{\mu_2})'\Sigma_2^{-1} \\ &\quad (\boldsymbol{x}-\boldsymbol{\mu_2})-(\boldsymbol{x}-\boldsymbol{\mu_1})'\Sigma_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu_1})] \\ &\quad \text{(neglecting the higher order terms)} \\ &= \frac{1}{2}log\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right)+\left(\frac{1}{2}+\frac{p}{2n}\right)\Delta^2. \end{aligned}$$

Thus

$$\begin{aligned} R_1 &: \frac{1}{2}.log\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right)+(\frac{1}{2}+\frac{p}{2n})\Delta^2>0. \\ i.e. \quad & \Delta^2 > \frac{n}{n+p}log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right). \quad (4) \end{aligned}$$

Since t distribution tends to Normal distribution as d.f. n tends to ∞, we see that as expected this rule tends to the QDA rule as n tends to ∞.

So, combining (**??**) and (**??**) we can define a decision rule as

$$\begin{aligned} x \in R_1 \quad & \text{if } \Delta^2 \geq c.log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) \\ x \in R_2 \quad & \text{otherwise .} \quad (5) \end{aligned}$$

for some constant $c \geq 0$ with $c = 0$ implying the classifier is Minimum Mahalanobis Distance(MMD) and $c = 1$ implying the classifier is Quadratic Discriminant Analysis. In this study our objective is to find a c, $c \neq 0$ and 1 which maximizes the accuracy of classification. For this purpose we proceed as follows. First using the training data set for the two classes, we estimate the value of $log\frac{|\Sigma_1|}{|\Sigma_2|} = d$, say. Next we compute the Δ^2 values for the all the observations in the training set of both the classes. Let $r_{1(1)},\ldots,r_{1(n_1)}$ denote the ordered Δ^2 values for class-1 and $r_{2(1)},\ldots,r_{2(n_2)}$ denote the Δ^2 values for class-2. Then we sort the Δ^2/d values for both the classes together. If $r_{2(n_2)} < r_{1(1)}$ then c can be chosen as any value in $[r_{2(n_2)}, r_{1(n_1)}]$. If $r_{2(n_2)} > r_{1(1)}$, determination of c needs a careful judgment. If for some $s \geq 1$, each of $r_{2(n_2-s+1)},\ldots,r_{2(n_2)}$ exceeds $r_{1(1)}$, treating each of these values as c, the misclassification error in the training set is calculated and finally a c is chosen having the least misclassification error. The resulting method is called Adaptive Nonparametric Discrminant Analysis (ANDA) from hereon.

To show the effectiveness of ANDA, we choose two simulation models. As Quadratic Discriminant Analysis is optimal when the class densities are multivariate normal, we experimented with this distribution first. In this case we expect that the chosen c (denoted by c^* from hereon) by this method will be close to 1, the optimal value corresponding to this distribution. We have chosen 4-variate normal distribution for the 2 classes with mean vectors and dispersion matrices as given below:

Table 1: Parameters for Simulated Normal

	Mean vector ($\boldsymbol{\mu}$)	Dispersion matrix ($\Sigma = ((\sigma_{ij}))$)
Class I	(0, 1, 2, 3)′	$\sigma_{ii} = 2\ \forall i, \quad \sigma_{ij} = 1\ \forall i \neq j$
Class II	(1, 0, 3, 2)′	$\sigma_{ii} = 4\ \forall i, \quad \sigma_{ij} = 0.5\ \forall i \neq j$

We have simulated 2000 observations from each class, and used 80% of the data for training and the remaining 20% as test cases. Table 2 presents in percentages the misclassification error (ME%) found in 8 such simulations :

Table 2: ME % for Simulated Normal data

c^*	Train ME% for			Test ME% for		
	c=0	c^*	c=1	c=0	c^*	c=1
0.845271	32.80	15.20	16.10	31.35	17.75	16.50
1.077770	33.50	15.00	15.30	31.95	17.10	17.20
1.169418	31.90	15.60	16.45	34.55	16.15	16.30
1.275173	32.85	15.20	15.75	32.35	17.25	16.65
1.104841	32.20	15.85	16.30	32.40	16.10	15.95
1.270853	33.35	15.05	15.60	32.45	17.30	16.50
1.052938	34.25	15.05	15.35	33.20	17.15	17.40
1.193218	31.90	15.30	15.75	33.15	17.05	16.80
Average	32.84	15.28	15.83	32.68	16.98	16.66

In the above table, misclassification errors are given for the different values of c for both the training and test sets. Hence, the errors corresponding to $c = 0$ refers to the MMD classifier, $c = 1$ refers to the QDA classifier, and c^* refers to the proposed method ANDA. As expected, QDA performs much better than MMD. It is encouraging to see that the selected c^* is very close to 1 in most of the cases and the error of ANDA is very close to that of the QDA, even less in some simulations.

Now let us consider the case when the class densities follow the multivariate Cauchy distribution (multivariate t-distribution with 1 degree of freedom). In this case, mean

and dispersion matrices do not exist for this distribution. We choose the centers of the two distributions as the same as mean vectors as before. The scale matrix for the first class is chosen to be the same as the dispersion matrix before but for the second class it is a diagonal matrix with the diagonal entries being all 4. We have again simulated 2000 observations from each class, and used 80% of the data for training and the remaining 20% as test cases.

Table 3 shows the results of 8 such simulations:

Table 3: ME % for Simulated Cauchy data

c^*	Train ME% for			Test ME% for		
	c=0	c^*	c=1	c=0	c^*	c=1
0.0026	13.81	13.13	49.56	12.00	10.50	49.75
0.0004	15.75	15.69	48.94	11.75	12.00	51.75
0.0003	14.56	14.44	50.25	13.75	13.75	46.50
0.0010	13.50	13.44	48.63	13.75	14.25	52.75
0.0004	12.31	12.19	49.69	12.25	12.50	47.75
0.0008	13.56	13.38	49.25	16.00	15.50	49.25
0.0062	16.13	15.69	48.44	20.25	19.50	50.75
0.1377	17.19	15.75	36.75	16.00	16.50	39.00
Average	14.60	14.21	47.69	14.47	14.31	48.44

As expected, MMD performs much better than QDA. It is encouraging to see that the selected c^* is very much close to 0 in all the cases except one, and the error of ANDA is very close to that of the MMD, yielding even less error in some simulations.

We now apply the method in two real datasets obtained from the Machine Learning Repository, UCI., viz. Diagnostic Wisconsin Breast Cancer dataset and MAGIC Gamma Telescope dataset, both having two classes. We randomly select 80% of the observations for training and use the remaining 20% as test cases. This process is repeated 10 times and we report the average misclassification error (in percentages) along with it's standard deviation (in parenthesis) below:

Table 4: ME % for two real datasets

Data	Train ME%(SD) for			Test ME%(SD) for		
	c=0	c^*	c=1	c=0	c^*	c=1
DWB Cancer	8.07 (1.87)	1.47 (0.66)	1.78 (0.77)	11.97 (2.73)	4.82 (1.22)	5 (1.26)
MG Telescope	43.21 (0.99)	19.87 (0.24)	21.18 (0.26)	42.70 (1.10)	19.82 (0.25)	21.10 (0.27)

From the table, we observe that in both cases, QDA is performing better than MMD. What is really interesting is that the proposed method even outperformed QDA in both the cases.

3. Multi-class problem:

Now we consider the case where there are m classes, $m > 2$. The multi-class classification problem refers to assigning each of the individuals into one of the m classes. As the two-class classification problem is much easier to handle, many authors propose to use two-class classifiers stepwise for multi-class classification. Our objective is to find a value of c to minimize the number of misclassification. We illustrate our procedure with three classes assuming n_i observations from Class i in the training set. Let the mean/location vectors and dispersion/scale matrices for Class i be denoted by $\boldsymbol{\mu_i}$ and Σ_i respectively, for $i \in \{1, 2, 3\}$. Let

$$\begin{aligned} \Delta^2_{ij}(\boldsymbol{x}) &= \frac{1}{2}[(\boldsymbol{x}-\boldsymbol{\mu_j})'\Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu_j}) \\ &\quad -(\boldsymbol{x}-\boldsymbol{\mu_i})'\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu_i})] \\ d_{ij} &= \frac{1}{2}.log(\frac{|\Sigma_j|}{|\Sigma_i|}) \\ u_{ij}(\boldsymbol{x}) &= \frac{\Delta^2_{ij}(\boldsymbol{x})}{d_{ij}}. \end{aligned}$$

For every ordered pair of classes $(i, j),\ \ i, j \in \{1, 2, 3\},\ i \neq j$, we calculate $u_{ij}(\boldsymbol{x})$s, for all training set observations $\boldsymbol{x}$ in Class i and Class j, order them and consider only those values of $u_{ij}(\boldsymbol{x})$ such that $u_{ij}(\boldsymbol{x}) \in [0, 1]$. Let the pool of such $u_{ij}(\boldsymbol{x}) \in [0, 1]$ varying $i, j \in \{1, 2, 3\},\ i \neq j$ be denoted by T. We consider each of these $u_{ij}(\boldsymbol{x}) \in T$ as c in turn to calculate the overall misclassification error. For each such $c \in T$, we identify the rightly classified Class 1 observations of the training set by comparing $u_{12}(\boldsymbol{x})$ and $u_{13}(\boldsymbol{x})$ with c separately. Let the set of rightly classified Class 1 observations from these two comparisons be denoted by R_{12} and R_{13} respectively. Then the number of misclassified Class 1 observations is $MC_1 = n_1 - |\{R_{12} \bigcap R_{13}\}|$. We repeat the same procedure to identify the misclassified Class 2 observations comparing $u_{21}(\boldsymbol{x})$ and $u_{23}(\boldsymbol{x})$ with c separately and calculate MC_2. Similarly the misclassified Class 3 observations are identified and the number of overall misclassification is $MC_1 + MC_2 + MC_3$. Such procedure is repeated for every $c \in T$ and the one with least number of overall misclassification is chosen to work with.

We applied the above procedure to some simulated datasets form Normal, t distribution and Cauchy distribution, and some real datasets. The Mean vectors ($\boldsymbol{\mu}$) and the Dispersion matrices (Σ) for *Normal distribution* are given below:

Table 5: Parameters for multiclass Simulated Normal

	$\boldsymbol{\mu}$	$\Sigma = ((\sigma_{ij}))$
Class I	(0, 1, 2, 3)$'$	$\sigma_{ii} = 2\ \forall i$, $\sigma_{ij} = 1\ \forall i \neq j$.
Class II	(1, 0, 3, 2)$'$	$\sigma_{ii} = 4\ \forall i$, $\sigma_{ij} = 0.5\ \forall i \neq j$.
Class III	(3, 0, 1, 4)$'$	$\sigma_{ii} = 1\ \forall i$, $\sigma_{ij} = \begin{cases} 0.5 & \text{if } \lvert i-j\rvert = 1, \\ 0 & \text{otherwise.} \end{cases}$

We simulate from 4 variate *t-distribution* with degrees of freedom 2, each class having the same mean vectors as simulated Normal distribution but different choices of

dispersion matrices as given below :

Table 6: Dispersion matrices for multiclass t-distribution

	Dispersion matrix ($\Sigma = ((\sigma_{ij}))$)
Class I	$\sigma_{ii} = 2\ \forall i,\quad \sigma_{ij} = \begin{cases} 1.5 & \text{if } \lvert i-j\rvert = 1, \\ 1 & \text{otherwise.} \end{cases}$
Class II	$\sigma_{ii} = 4\ \forall i,\quad \sigma_{ij} = \begin{cases} -1 & \text{if } \lvert i-j\rvert = 1, \\ 0 & \text{otherwise.} \end{cases}$
Class III	$\sigma_{ii} = 3\ \forall i,\quad \sigma_{ij} = \begin{cases} -1 & \text{if } \lvert i-j\rvert = 1, \\ 0 & \text{otherwise.} \end{cases}$

We simulate from 8 variate *Cauchy distribution* with location vectors same as the mean vectors of the simulated Normal distribution and the scale matrices as given below:

Table 7: Scale matrices for multiclass Simulated Cauchy

	Scale matrix ($\Sigma = ((\sigma_{ij}))$)
Class I	$\sigma_{ii} = 2\ \forall i,\quad \sigma_{ij} = \begin{cases} 1.5 & \text{if } \lvert i-j\rvert = 1, \\ 1 & \text{otherwise.} \end{cases}$
Class II	$\sigma_{ii} = 4\ \forall i,\quad \sigma_{ij} = \begin{cases} -1 & \text{if } \lvert i-j\rvert = 1, \\ 0 & \text{otherwise.} \end{cases}$
Class III	$\sigma_{ii} = 3\ \forall i,\quad \sigma_{ij} = \begin{cases} -1 & \text{if } \lvert i-j\rvert = 1, \\ 0 & \text{otherwise.} \end{cases}$

Four real datasets viz. *New-Thyroid data*, *Vowel data*, *Ecoli data*, *PB Classification* are obtained from Machine Learning Repository, UCI.

The results of classification using the proposed method(ANDA) are shown below:

Table 8: ME% in multi class classification

Data	# of classes	c^*	ME% Train for			ME% Test for		
			c=0	c^*	c=1	c=0	c^*	c=1
Normal	3	0.9858	39	14.27	14.33	38.07	15.53	15.6
t (df 2)	3	0.0533	17.8	17.13	39.07	19.4	19.33	40.2
Cauchy	3	0.1286	24.4	22.47	36.27	25.47	24.73	35.2
Thyroid	3	0.4921	5.56	2.78	4.63	6.54	3.74	3.74
Ecoli	4	0.3174	7.74	6.55	6.55	9.52	8.33	7.14
Vowel	10	0.4993	21.6	20.71	21.6	19.82	21.02	21.02
PB	5	0.9034	8.66	5.07	5.23	9.64	5.63	5.66

From the above table we see that, in the case of simulated data, the misclassification error of ANDA is close to the one with that value of c we should expect theoretically, i.e. $c = 1$ for Normal and $c = 0$ for t and Cauchy. In some examples, QDA (corresponding to $c = 1$) performs better than MMD (corresponding to $c = 0$), while it's exactly the opposite in the other examples. However, ANDA has the lowest misclassification error except for the Vowel data and Ecoli data. It also figures out the optimal choice of the threshold-value c very effectively.

4. Comparison of ANDA with Standard Multivariate Classification Methods:

We compare the performance of ANDA with other standard classification methods existing in the literature such as: Tree, Neural Network (NN), Support Vector Machine (SVM) with several different kernels, k Nearest Neighborhood (k-NN), Naive-Bayes and Discriminant Analysis to find out whether the proposed methods are competitive or not. The reader is referred to Hastie et al.[**?**] for details of these methods. In our comparison procedure, we randomly partition each data set into a training data set and test data set containing approximately 80% and 20% of the whole data respectively. Here we use the four real datasets mentioned above, viz. Vowel data, New-Thyroid data, Ecoli data and PB classification data. The results are given in the following table:

Table 9: Comparison of ANDA with other methods

Method	ME%			
	Vowel	New-Thyroid	Ecoli	PB
Standard Tree	24.02	13.08	7.14	5.76
Neural network	23.12	2.80	12.50	4.20
SVM	20.12	5.61	8.33	6.22
k-NN	21.62	7.48	8.33	4.66
Naive Bayes	25.23	3.74	8.93	9.60
ANDA	21.02	3.74	8.33	5.63

As it is evident from the above table that ANDA is very competitive with the other established sophisticated classifiers. It's misclassification error is very close to the minimum. No other method is uniformly better than ANDA which is very simple with very low computational cost.

5. Conclusion

This paper attempts to generalize the QDA and MMD classifiers and proposes a new method ANDA which is very simple, purely nonparametric and computationally cost-effective. This flexible method adapts to a particular dataset by choosing an optimal threshold value so that it can match the performance of QDA when it is optimal. It also produces very good results when QDA fails miserably. The experiments suggest that ANDA is very competitive with other established nonparametric methods.

References

[1] T. W. Anderson *An Introduction to Multivariate Statistical Analysis*, New York: John Wiley and Sons, 1958.

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer-Verlag, 2001.

Inference for the Preference of Program Genre Using Audience Measurement Information

Sang-Yun Lee[1], Sang-Taick Park[1], Jin-Woo Hong[1], Chuho YI[2], and Seungdo Jeong[3]
[1] Electronics & Telecommunications Research Institute, Smart TV System Research Team, 218, Gajeongno, Yuseong-gu, Daejeon, 305-350, Korea
[2]Hanyang University, Haengdang-dong Sungdong-gu, Seoul 133-791, Korea
[3]Department of Information and Communication Engineering, Hanyang Cyber University, Korea

Abstract - *In this paper, we propose inference algorithm for the program genre based on audience measurement using the attention function. We eliminate the short-term time spent in selecting the channel in calculating the preference for the genre. To evaluate the performance of the proposed algorithm, we use the former 5 month's data in training and calculating the probability of preference and compare them with the latter 1 month's data.*

Keywords: audience measurement, Bayesian, attention function, beta distribution

1 Introduction

The preference for the genre is used in producing a TV program or scheduling an advertisement. Recently, the interests in targeting services grow increased as interactive services become possible through a smart TV. To infer the preference of genre audience measurement information such as usage history, retrieval or purchasing history can be used.

In this paper, we propose inference algorithm for the preference of the genre through analyzing audience measurement information. Wonneberger suggested modeling of viewing behavior with channel changing and program selection. In that modeling, he defined it as modeling sequence of viewing [1]. But we will advance the accuracy by introducing attention function. To evaluate the performance of the proposed algorithm, we will train and calculate the probability of preference of the genre using the former 5 month's data and will compare them with the latter 1 month's data. The remainder of this paper is organized as follows. In section 2, we introduce the inference algorithm for the genre. In section 3, we describe its test result. Finally, we summarize and conclude this paper in section 4.

2 Preference Inference Algorithm

2.1 Attention Function

According to the length of the viewing time we can determine whether an audience has interests in that program or not. Therefore, in case an audience views a program only within a certain amount of minutes, the time spent in viewing should be eliminated in calculating the preference of the genre, and should be given more weights in reverse case. We introduce the attention function in order to reflect the degree of interests with program time. The attention function needs starting and ending time of the program to calculate the length of the program and the time spent in viewing the program. We approximates the attention function with beta distribution function [2] described by

$$f_{attention}(v_i;\alpha,\beta) \quad (1)$$

where α and β determines the form of beta distribution, we set them with 4 and 1.5 respectively.

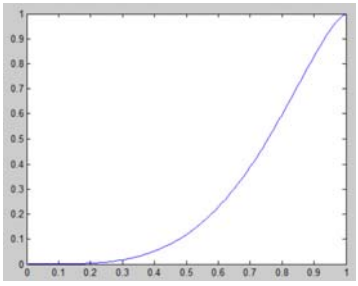


Fig. 1. Attention function using beta distribution

Fig. 1. represents the attention function. The x-axis denotes the ratio of the length of the program time and the length of the viewing time.For example, if you viewed the program of 50 minutes length during 5 minutes, it wouldn't be considered that you have interests in that program. In reverse, if you viewed that program during 40 minutes, it would be considered that you have many interests in that program, so the weight for the preference will be higher.

2.2 Calculation of the Preference of the Genre

We will represent the value of preference for the genre with probability. So that, we will apply Bayesian model that can reflect previous viewing behavior [3], which is calculated by

Table 1. Inference result of the preference for the genre

Item	A's Genre Preference (5 months)	A's Observed viewing time (1 month)	B's Genre Preference (5 months)	B's Observed viewing time (1 month)
First	Drama	Drama	Drama	Drama
Second	Entertainment_Etc	Entertainment_Etc	News	Entertainment_Etc
Third	Talkshow	Talkshow	Documentary	News
Fourth	News	Entertainment_Total	Entertainment_Etc	Documentary
Fifth	Entertainment_Total	News	Life	Talkshow

$$P(g_i \mid V) = \frac{P(g_i)P(V \mid g_i)}{P(V)} \quad (2)$$

where V is the audience's viewing data, g_i is the i th genre and $P(g_i \mid V)$ is the posterior probability of i th genre calculated by observed data. $P(g_i)$ represents the prior probability of the genre, and it means the occurring probability of the each genre. We assume that the probability of the each genre is uniformly distributed. $P(V \mid g_i)$ denotes the likelihood between the observed data and the corresponding genre. Therefore, the value of $P(V)$ can be eliminated and the likelihood can be calculated again according to the observed data by

$$\propto P(g_i)\prod_{i=1}^{|V|} P(v_i \mid g_i) \quad (3)$$

The calculated value from (3) can be under-flow by multiplying decimal points continuously, so we take a logarithm of it as (4)

$$\propto \log P(g_i) + \log \sum_{i=1}^{|V|} f_{attention}(v_i; \alpha, \beta) \quad (4)$$

Then, we apply the attention function to calculate the likelihood using the observed data

$$v_i = \frac{\text{watching length}}{\text{program length}} \quad (5)$$

where v_i is used as the input value of observed data and it is obtained by dividing the length of viewing time to the length of the program time.

3 Experimental Results

To train and calculate the audience's preference for the genre we used two user's actual viewing behavior data measured and provided by AGB Nielson Company. Those data are gathered from 1 April 2011 to 30 September 2011. To evaluate the performance of the proposed algorithm, we calculated the probability of the preference for the genre using the former 5 month's data and compared them with the latter 1 month's data. The Kobaco provides 47 medium genre scheme such as music (show), documentary, news, entertainment_etc, talkshow, and information, and we used them for genre classifying. In experiment, we compared the top fifth ranks among them.

From table 1 we can see that A's rank of preference matches from first to third, and the fourth and the fifth is reversed while B's first rank matches each other, but other ranks do not match. The proposed algorithm in this paper had inferred the preference of the genre accurately for A, but it had inferred it inaccurately for B. This is because B's latter 1 month's viewing behavior was different considerably from the former 5 month's that.

4 Conclusion

In this paper, we proposed inference algorithm of the preference for the genre using attention function, and calculated the preference based on Bayesian model. We eliminated the short-term time spent in selecting the channel in calculating the preference because the more an audience views a program the higher the weight for the preference increases.

The experimental results show that the proposed algorithm can infer user's preference well comparatively and we will develop a recommending system for a TV program or an advertisement based on the proposed method.

5 References

[1] Wonneberger, K. Schoenbach, and L. V. Meurs, "Dynamics of Individual Television Viewing Behavior: Models, Empirical Evidence, and a Research Program," Communication Studies, 60(3), pp. 235-252, Jun. 2009.

[2] C. G. Verdugo Lazo and P. N. Rathie. "On the entropy of continuous probability distributions," IEEE Trans. Inf. Theory, pp. 120–122, 1978.

[3] H. John and P. Langley, "Estimating Continuous Distributions in Bayesian Classifiers," Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, pp. 338-345, 1995.

Acknolegement

This work was supported by the ETRI R&D Program of KCC(Korea Communications Commission), Korea [11921-03001, "Development of Beyond Smart TV Technology"].

SESSION

NOVEL TECHNIQUES AND ALGORITHMS

Chair(s)

TBA

Finding both Aggregate Nearest Positive and Farthest Negative Neighbors*

I-Fang Su[1], Yuan-Ko Huang[2], Yu-Chi Chung[3,], and I-Ting Shen[4]**
[1]Dept. of IM,Fotech, Kaohsiung, Taiwan [2]Dept. of IC, KYU, Kaohsiung, Taiwan
[3]Dept. of CSIE, CJCU, Tainan, Taiwan [4]Dept. of CSIE, NCKU, Tainan, Taiwan

Abstract— *Recently, researchers use the aggregate nearest neighbor (ANN) search for users at different locations (query points) that want to find one restaurant (data point), which leads to the minimum sum of distances where they have to travel in order to meet. Users can also use aggregate farthest neighbor (AFN) search to find one of the building locations of a new hotel (query points) so as to maximize the aggregate distances to all the other existing hotels(data points) for reducing competition. These works mainly focus on finding either the aggregate nearest neighbors or the aggregate farthest neighbors. In reality, users not only have queries of aggregate nearest neighbors but also have queries of the aggregate farthest neighbors. He needs to make a decision through both aggregate nearest neighbors such as finding the objects that user prefer from the a set of data points and aggregate farthest neighbors which retrieves the objects that user dislike from another set of data points. In order to verify these two sets of query points, we named the objects which user prefer the Positive query set, and the objects that user dislike the Negative query set. Motivated by these observations, we propose a novel query by combining the aggregate nearest positive neighbor search and the aggregate farthest negative neighbor search together meaningfully. We name this query the Aggregate Nearest Positive and Farthest Negative Neighbors (ANPFNN) query. In this paper, we propose a round-robin algorithm to retrieve the first aggregate nearest positive and farthest negative neighbors. Further, we use a pruning rule to efficiently filter out the answers. Our extensive evaluation results validate the effectiveness and efficiency of our algorithm on uniform distributed clustered data.*

Keywords: aggregate nearest positive, farthest negative neighbor search, nearest neighbor, dominate

1. Introduction

Nearest neighbor queries (NN queries) [1], [2], [3], [4] have been widely discussed in recent years. Given an n-dimensional data set D and a query point q, an NN querie finds the data point d (d∈D) which is nearest to q. For example, in Figure 1, a point data set P (P=$\{p_1, p_2, p_3, p_4, p_5\}$) represents the locations of parking lots, and a point q is the current position of the user. If a user issues a query for finding the location of the nearest parking lot to a server, the server returns p_2 as it is closest to q within the data point set P.

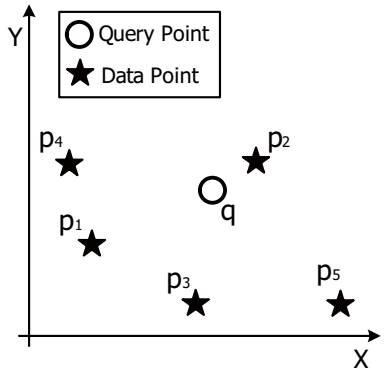


Fig. 1: Example of an Nearest Neighbor search.

Recently, aggregate nearest neighbor (ANN) query[5], [6] had been proposed. ANN queries are a variant of NN queries. The conventional NN query is to find a data point p_i that is close to a given query point q, where p_i is within a point data set P=$\{p_1, p_2, p_3, \cdots\}$. The ANN query problem involves a point data set P=$\{p_1, p_2, p_3, \cdots\}$, a point query set, $Q = \{q_1, q_2, q_3, \cdots\}$, and a given function F such as **MIN**, **MAX**, and **SUM**. An ANN query is to find a point p_i within P that has the shortest function distance to all the query points. For example in Figure 2, suppose a university holds an international conference, accommodations are arranged for the attendees at several hotels ($Q = \{H_1, H_2, \cdots, H_7\}$) near the university. The conference is looking for a venue to hold a banquet from given restaurants $P = \{R_1, R_2, R_3, R_4\}$. To minimize the total transportation costs of all the attendees (supposing that the transportation rates are the same), the host must find one restaurant from P that has the shortest total distance to Q. The Table 5 depicts

*This work is supported by National Science Council of Taiwan (R.O.C.) under Grants NSC 100-2221-E-268 -007, NSC100-2221-E-309-011, NSC 100-2221-E-244 -018.
**Corresponding Author

the shortest distances from the restaurants to each of the hotels. To minimize transportation costs, the host chooses to hold the banquet at R_2. Because of all the restaurants, the total distance between R_2 and all of the hotels is the shortest(14.2).

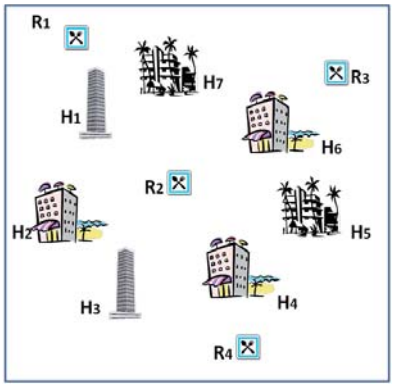


Fig. 2: Example of an Aggregate Nearest Neighbor search.

In addition to the **SUM** function, **MAX** and **MIN** are also employed as set functions in ANN queries. We use Figure 2 to explain the varying meanings of the other functions in ANN queries. If **MAX** is adopted as the set function, the ANN query returns R_2. The reason is that the maximum distance of R_2 to all hotels is the shortest among that of other restaurants. This ANN query shows that selecting R_2 as the venue for the banquet minimizes the time attendees spend getting there (supposing that all of the attendees move at the same speed). If **MIN** is adopted as the set function, the ANN query returns R_4 since the minimum distance of R_4 to all hotels is the shortest among that of other restaurants. The primary objective of this query shortens the vacancy time of the venue (because R_4 can be reached the most quickly).

Another variation of NN query is the aggregate farthest neighbor (AFN) query [7]. Given a point data set P, a point query set Q, and a given function F, a data point p ($p \in P$) is derived with the maximum distance function with query set Q. For example, suppose a financial group wishes to build a resort at popular tourist destination, they must find a location p among the potential areas P as far as possible from all the other hotels Q. Similar to ANN queries, the AFN queries also include **MAX**, **MIN**, and **SUM** as their aggregate function.

From the above observation, we found that either the ANN queries or the AFN query only gives the one-sided information to users. In reality, users may need to facilitate the exploration of a data space and make an objective decision. For instance, people usually hope for a house with high functionality in the surroundings but away from dangerous or noisy locations while they are purchasing houses. If we present these characteristics in the form of options, home buyers may hope for the presence of (1) schools, (2) MRT stations, and (3) parks within a 500-meter radius. At the same time, hazardous locations such as (1) power plants, (2) gas stations, and (3) industrial zones are not desired within neighboring areas. For example in Figure 4, suppose that X and Y are distance coordinates, d_i is the house for sale in the area where i = 1 to 4, P=$\{p_1, p_2, p_3\}$ is the query set of locations that the user wishes to be closer to (such as parks, schools, and MRT stations), and N=$\{n_1, n_2, n_3\}$ is the query set of locations that the user wishes to be farther away from (such as gas stations, power plants, and chemical plants). In order to verify these two types of query sets, we use the *Positive query set* and *Negative query set* to represent the locations that user wish to be closer to and farther away from, respectively. In this example, the user may select one house from the given data points that is closer to the positive query set and is simultaneously farther away from the negative data points. As seen in Figure 4, under the preference conditions of the user, d_2 and d_3 would be the two recommanded houses to the user's needs. The reason is that the maximum distance of d_2 to the positive query set is the shortest and the minimum distance of d_2 to the negative query set is the farthest. In addition to d_2 is a recommend house for users, d_3 is also an option for users. Although the maximum distance of d_3 to the positive query set is not the shortest, the minimum distance of it to the negative query set is the farthest. That is to say, d_3 is the house which farther away from the locations user dislike. And d_2 and d_3 are both answers of this given query.

Apart from the above example, there are a large number of similar queries in real life. From the above observations, we believe that this kind of query is important for users to make a decision. However, previous studies are unable to apply existing methods to this type of query. Although ANN and AFN queries are related to this type of query, they only consider either aggregate nearest or farthest neighbor search. The answers provided by these two algorithms can only satisfy one single user demand, and cannot simultaneously find the result from this two different query sets. Besides, there may not be only one result which meets the demands of users and ANN as well as AFN queries may only satisfy a portion of the query conditions. Therefore, we formally define the problem of finding the Aggregate Nearest Positive and Farthest Negative Neighbors query (ANPFNN) in this paper. In this paper, we propose an algorithm to execute the ANPFNN query and use a pruning rule to efficiently filter out the answers so as to reduce the computation cost. We also apply a dominance test [8] in ANPFNN in order to retrieve all significant objects for users to make a good decision. Our extensive evaluation results validate the effectiveness

Distance (km)	H_1	H_2	H_3	H_4	H_5	H_6	H_7	Total
R_1	2	4	5.5	7	5.5	4	2	30
R_2	2	1.5	2	1.5	3	2	2.2	14.2
R_3	5.5	7	6	4.5	2.5	1.5	3.5	30.5
R_4	6	4.5	2.2	1	2.5	3	7	26.2

Fig. 3: The distances between restaurants and hotels of Figure 2.

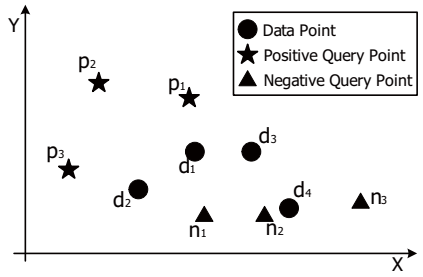


Fig. 4: Example of an aggregate nearest positive and farthest negative neighbors search.

and efficiency of our algorithm on both uniform distributed and clustered data.

The rest of the article is organized as follows. Section 2 reviews related work on ANN and AFN searches. Section 3 provides preliminaries on ANPFNN query and some distance metrics used in this paper. Section 4 presents our main algorithm for processing ANPFNN queries efficiently. Results of our experimental study are reported in Section 5. Finally, Section 6 concludes the paper.

2. Related Work

In this section, we briefly review previous work related to ANPFNN queries. Papadias et al. [5] proposed three algorithms, named *Multiple Query Method*(MQM), *Single Point Method*(SPM), and *Minimum Bounding Method*(MBM) for processing aggregate nearest neighbor searches. In [7], Yuan Gao et al. propose *minimum bounding*(MB) and *best first*(BF) algorithms for processing aggregate farthest neighbor queries. The main idea of the first algorithm exends the idea of MBM in ANN to MB by using a threshold to filter out the possible results, and then efficiently retrieve the answers. Instinctively, ANPFNN can be disassembled into an ANN problem and an AFN problem. However, adopting this existing method for ANPFNN problems presents the following issues. In order to accelerate query speed, ANN and AFN both employ R-tree to index the query set and data points. Hence, the root nodes of R-trees are visit frequently while processing query for finding points nearest to the positive query set and those farthest from the negative data set. Subsequently, this step incurs a considerable amount of computation costs. And many redundant query results appear during processing. Moreover, the finding results may only fit for one-sided query and a further process is required for retrieving the final results. Therefore, we apply a different pruning technique that can quickly filter impossible answers to further reduce the required computation costs involved in processing queries.

3. Preliminaries

In this section, we first give the definitions of the aggregation distance, the dominance relation, the aggregate nearest positive as well as farthest negative neighbors, and formally define the aggregate nearest positive and farthest negative neighbors (ANPFNN) query. We further describe the underlying indexing structure of our algorithm in this section.

3.1 Definitions

Given a set of data points D= $\{d_1, d_2, \cdots, d_i\}$, a set of positive query points P=$\{p_1, p_2, \cdots, p_x\}$, and a set of negative query points N=$\{n_1, n_2, \cdots, n_y\}$,the aggregation distance (D_{agg}) is defined in Definition 1.Based on Definition 1, the ANPFNN query is formulated in Definition 2.

Definition 3.1: **Aggregation distance (**$AggD$**)**
Given a data point d, the aggregation distance between d and a query set P is $AggD(d, P) = F_{j=1}^{x} \|d, p_j\|$, where F can be SUM, MAX or MIN.
In this paper, we use $Fsum$, $Fmax$, and$Fmin$ to represent the respective aggregation function.

Definition 3.2: **Nearest Positive distance**
Given a set of data points d, and a set of positive query points P=$\{p_1, p_2, \cdots, p_x\}$, the nearest positive distance is to find the maximum $AggD(d, P)$ of d to the set of positive query points P. Nearest Positive distance = $Fmax_{j=1}^{x} \|d, p_j\|$.

Definition 3.3: **Farthest Negative distance**
Given a data points d, and a set of negative query points N=$\{n_1, n_2, \cdots, n_y\}$, the farthest negative distance is to find the minimum $AggD(d, N)$ of d to the set of negative data points N. Farthest Negative distance =$Fmin_{j=1}^{y} \|d, n_j\|$.

We can then use the above two distances to define the aggreate nearest positive and farthest negative distances.

Definition 3.4: **Aggregate Nearest Positive distance,** $PAgg$
Given a set of data points D= $\{d_1, d_2, \cdots, d_i\}$, and a set of positive query points P=$\{p_1, p_2, \cdots, p_x\}$, the aggregate

	Positive Query Set			Negative Query Set		
Distance (km)	p_1	p_2	p_3	n_1	n_2	n_3
d_1	12	28	32	16	32	45
d_2	25	28	18	20	38	57
d_3	21	41	47	26	25	31
d_4	38	5	58	28	11	17

Fig. 5: The distances among data points, positive and negative query sets of Figure 4.

nearest positive distance is to find one data point d which has the minimum $AggD(d, P)$ among all data points in D to the set of positive query points P. $PAgg$ = MIN$\{AggD(d_t, P)\}$ = MIN$_{t=1}^{t=i}\{Fmax_{j=1}^{x} \|d_t, p_j\|\}$.

Definition 3.5: **Aggregate Farthest Negative distance,** $NAgg$

Given a set of data points D= $\{d_1, d_2, \cdots, d_i\}$, and a set of negative query points N=$\{n_1, n_2, \cdots, n_y\}$, the aggregate farthest negative distance is to find a data point d which has the maximum $AggD(d, N)$ among all data points in D to the set of negative data points N. $NAgg$ = MAX$\{AggD(d_t, N)\}$ = MAX$_{t=1}^{t=i}\{Fmin_{j=1}^{y} \|d_t, n_j\|\}$.

Definition 3.6: **dominate ($\succ$)**

A data point d_i dominates another data point d_j on both $PAgg$ and $NAgg$ if and only if $PAgg$ of d_i is less than that of d_j and $NAgg$ of d_i is large than that of d_j.

Definition 3.7: **Aggregate Nearest Positive and Farthest Negative Neighbor (ANPFNN) queries**

Given a set of data points D= $\{d_1, d_2, \cdots, d_i\}$, a set of positive query points P=$\{p_1, p_2, \cdots, p_x\}$, and a set of negative query points N=$\{n_1, n_2, \cdots, n_y\}$, ANPFNN is to find a data point d ($d \in D$) where d is not dominated by any other data point of D in both $PAgg$ and $NAgg$.

3.2 Indexing Structure

To reduce I/O and computation cost of processing ANPFNN query, a well underlying indexing structure is indispensable. In this paper, we use R-tree [9] as our indexing structure. In a R-tree, objects are recursively grouped in a bottom-up manner according to their locations. For instance, Figure 6(a) gives a two-dimensional example where eight data points p_1 to p_8. The corresponding R-tree is showing in Figure 6(b). Each entry of a leaf node of a R-tree has the structure ($o._{ptr}$,($o._x, o._y$)), where $o._{ptr}$ is a pointer to the actual data point in the database, and $o._x$ and $o._y$ represent the X and Y coordinate of data point o, respectively. Each entry of an internal node has the structure (MBR_E, (x_l, y_d, x_r, y_u), $E._{ptr}$), where MBR_E is the minimum bounding rectangle (MBR) that encloses all the data points in the child node E of this internal node, (x_l, y_d, x_r, y_u) represent the lower bound and upper bound of node E in X and Y coordinates respectively, and $E._{ptr}$ is a pointer to node E. In our paper, we use three R$_P$-tree, R$_N$-tree, and R$_D$-tree to represent the positive query set, the negative query set, and the given data set.

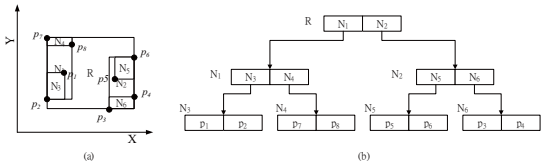


Fig. 6: Representations of entries in the R-tree

4. Aggregate Nearest Positive and Farthest Negative Neighbor Search Algorithm

A naive solution for processing a ANPFNN query is to scan all points and enumerate the $PAgg$ as well as $NAgg$ of each data point to the positive and negative query sets, respectively. Finally, applying the dominance test to each data point in both $PAgg$ and $NAgg$ to retrieve the final answers of ANPFNN. However, the time complexity of calculating dominance relationship, the $PAgg$ and the $NAgg$ of each data points is time consuming. Thus, we propose an ANPFNN algorithm which highly reduces the computation complexity. The ANPFNN algorithm consists of two main phases: (1)the filtering phase and (2)the refinement phase.

4.1 Filtering Phase

We first employing a branch-and-bound traversal on R$_P$-tree to find the data point d_i in D which d_i has the minimum $PAgg$. Then, the branch-and-bound traversal is also applied on R$_N$-tree to find the data point d_j in D which d_j has the maximum $NAgg$. This phase retrieves data points from R$_P$-tre and R$_N$-tree in a round-robin fashion until a data point d_t shows up in both R$_P$-tree and R$_N$-tree.

The way of finding the minimum $PAgg$ and maximum $NAgg$ are consider by Lu [10]. While traversing the nodes of the R$_P$-tree, a heap is applied to keep the MinDist(d, P) in an descending order where the MinDist(d, P) is the minimum possible distance from d to an object in R$_P$-tree. Then, we retrieve the first entry of the heap for further processing. This entry should has the maximum MinDist and it may contain the final result. If the entry is a data point, it

should be the possible answer for the minimum $PAgg$, and we can continue to the second step of the filtering phase. Otherwise the entry is repeatedly decomposed for finding the data point of this internal node. While traversing the nodes of the R_N-tree, we use another heap to keep the MaxDist(d, N) in an ascending order where the MaxDist(d, N) is the maximum possible distance from d to an object in R_N-tree. Follow the same procedure as finding the minimum $PAgg$, the first entry of the heap is retrieved for further processing. If the entry is a data point, it should be the possible answer for the maximum $NAgg$, and we can continue to the second step of the filtering phase. Otherwise the entry is repeatedly decomposed for finding the data point of this internal node.

4.2 Refinement Phase

When the filtering phase is terminated, many results are retrieved. However, these results may not meet both request of users such as the retrieved data point is near the positive query points but also near the negative query points). They may only be outperformed in one condition. Thus, we have to provide users the results which no other data can beat it in both conditions. We consider the B-tree [8] indexing method for finding the final results for users. We use two indexing for aggregate nearest positive and farthest negative distance and sort the retrieved data points according to the their $PAgg$ descendingly and $NAgg$ ascendingly. Then, scan through the whole index simultaneously to find the first match results d_t. Any result which has not listed after d_t is definitely not part of the answer because it is dominated by d_t. Meanwhile, the points listed before should be the final results for users to make decision. Due to space limitations, we refer readers to [11] for more detail examples about these two phases.

5. Performance

In this section we evaluate the efficiency of the proposed algorithm to the naive algorithm. The performance is measured by the average CPU time. The algorithms were developed in C++ and executed on a PC with a Intel i7 CPU of 2.8GHz. The default number of positive objects, negative objects and query points are all 110k, and we also vary the range of these three types of points from 10k to 210k. The coordinates of the objects are uniformly normalized in the domain $[0, 10000]^2$.

We compare the ANPFNN with naive algorithm under different number of negative objects, positive objects, and data points in Figure 7, 8, 9 respectively. Under these three circumstances, the average CPU time of ANPFNN increases smoothly, however, that of Naive algorithm increases dramatically. The performance results show that ANPFNN algorithm outperform the naive algorithm on the different object distributions. Due to space limitations, we show and the detail experimental evaluations are listed in [11].

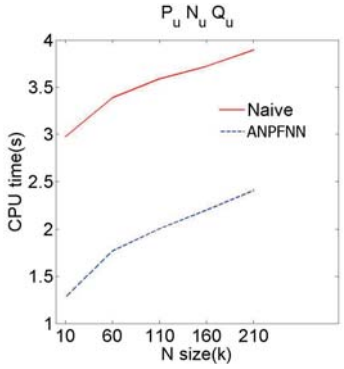


Fig. 7: The average CPU time under the number of negative query size.

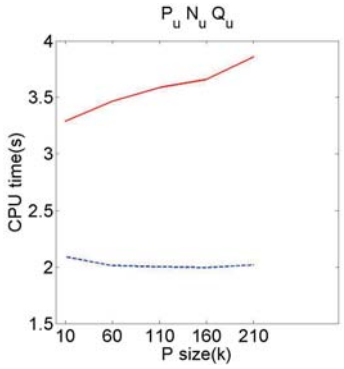


Fig. 8: The average CPU time under the number of Positive query size.

6. Conclusion

In this paper, we proposed the design and implementation of an algorithm for processing aggregated nearest positive and farthest negative neighbor queries in spatial networks. Our design applies the filtering phase to efficiently filter out all possible results and retrieve the final answers by using a refinement phase. Our performance study showed that this design exhibits a superior performance in terms of computation cost. The potential of ANPFNN query has not been fully exploited yet. Currently, we are extending the capability of this design to deal with query in a road network. An efficient query processing technique for processing ANPFNN query in this type of network is also under designed.

References

[1] B. Cui, B. C. Ooi, J. Su, and K.-L. Tan, "Indexing high-dimensional data for efficient in-memory similarity search," *ACM Transcations on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 339–353, 2005.

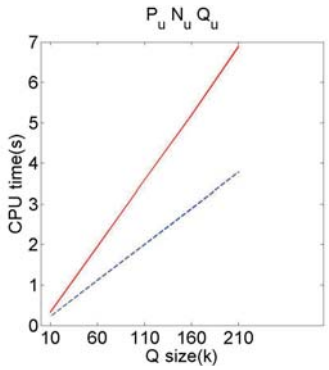


Fig. 9: The average CPU time under the number of data point size.

[2] L. Hong, B. C. Ooi, H. T. Shen, and X. Xue, "Hierarchical indexing structure for efficient similarity search in video retrieval," *ACM Transcations on Knowledge and Data Engineering*, vol. 18, no. 11, pp. 1544–1559, 2006.

[3] B. Zheng, J. Xu, W. chien Lee, and D. L. Lee, "Energy conserving air indexes for nearest neighbor search," in *Proceedings of the 9th International Conference on Extending Database Technology*, 2004, pp. 48–66.

[4] X. Xiong, M. F. Mokbel, and W. G. Afre, "Sea-cnn: Scalable processing of continuous k-nearest neighbor queries in spatio-temporal databases," in *Proceedings of International Conference on Data Engineering*, 2005, pp. 643–654.

[5] D. Papadias, Y. Tao, K. Mouratidis, and C. K. Hui, "Aggregate nearest neighbor queries in spatial databases," *ACM Transcations on Database Systems*, vol. 30, no. 2, pp. 529–576, 2005.

[6] M. L. Yiu, N. Mamoulis, and D. Papadias, "Aggregate nearest neighbor queries in road networks," *ACM Transcations on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 820–833, 2005.

[7] Y. Gao, L. Shou, K. Chen, and G. Chen, "Aggregate farthest-neighbor queries over spatial data," in *International conference on Database Systems for Advanced Applications, DASFAA*, 2011, pp. 149–163.

[8] S. Borzsonyi, D. Kossmann, and K. Stocker, "The skyline operator," in *Proceedings of the 17th International Conference on Data Engineering, ICDE 2001*, 2001, pp. 421–430.

[9] A. Guttman, "R-trees: A dynamic index structure for spatial searching," in *Proceedings of the 1984 ACM SIGMOD international conference on Management of data*, 1984, pp. 47–57.

[10] H. Lu, "On computing farthest dominated locations," *Journals and Magazines*, vol. 23, no. 6, pp. 928–941, 2011.

[11] I.-F. Su, Y.-K. Huang, Y.-C. Chung, and I.-T. Shen, "Anpfnn query," 2012, http://140.116.247.159/anpfnn.docx.

Reference Model for the Design of Compatible ERP-Modules in the Machinery and Equipment Industry

G. Schuh[1], **S. Cuber**[1]
[1]Institute for Industrial Management (FIR) at RWTH Aachen University, Aachen, Germany

Abstract -*Companies today face the challenge to encounter increasing market dynamics with the help of rather static IT-systems. Especially the monolithic ERP-systems require high cost and time efforts for adaptation and lead to the use of additional, non-integrated applications in order to serve process requirements in producing companies. This leads to a fragmented IT-system-landscape that prohibits a real-time, valid representation of the as-is situation and that is barely adaptable. This significantly limits a company's ability to detect deviations and to govern the order management under dynamic conditions. The paper at hand focuses on a comprehensive research project dealing with modularization of ERP systems into compatible modules that can be configured to a whole system within the approach of a service oriented architecture. An integrated information management model regarding the IT-support for order management processes within the engineer-to-order production is developed. Combining process activities with the perspective of data-flows and IT-functionalities, the model builds the bases for the clustering of IT-functionalities to ERP-modules and the design of their respective interfaces.*

Keywords: Manufacturing Planning, Operations and Control, Process Management, Information Management, ERP-Systems

1 Problem situation

The producing industry in high-wage countries today is facing the requirement to adapt in a fast and flexible way to increasingly volatile market conditions. A differentiation from competition is no longer only achieved by the manufactured products but more and more by a dynamic, changeable design of the business processes of order fulfillment. These have to be continuously monitored and adapted to the customers' requirements [1], [2]. The implementation of agile and at the same hand stable processes are the main goals within the design of production systems [3].

In order to be able to conduct an efficient and dynamic (re-)organisation of order management processes each company is dependent on IT-systems [4]. According to the described background this requires today more than ever flexible IT-structures that are easy to adapt and perfectly fit and support the companies' processes. However this demand does not comply with today's IT-system landscapes in companies. Missing data- and information-integration as well as missing interoperability between the single applications inside and between companies inhibit a fluent, real-time information flow.

A main reason can be found in today's ERP-systems which build the backbone of companies' production planning and control. Regarding their usually complex, monolithic layout, they are not designed to support a fast and flexible adjustment to business conditions and require high efforts in time and costs for adaptation. Monolithic in that sense means that even though many ERP-systems already consist of several modules, the functionality of these modules as well as their interfaces vary widely between different ERP-system-providers (cf. Figure 1) [5]. Furthermore in practice these systems are highly customized at each single company. A combination respectively an exchange of these modules of different providers according to their individual strength and the companies' needs remains nearly impossible. The single system consequently is monolithic in the way that it is not designed for a provider-spanning configuration.

This situation has led to the utilization of additional software and stand-alone solutions in order to meet customer- and process-demands [6], [7]. The result is a heterogeneous IT-landscape, which causes high costs for the support and maintenance of all systems, does not allow a consistent monitoring and controlling of the as-is situation and handicaps or inhibits to foresee the consequences of an adjustment of the systems on the overall system landscape and the operational process performance.

However this heterogeneous and mostly inconsistent IT-system-landscape hardly enables an effective and efficient planning, operation and control of order management processes in a dynamic business environment. The distributed, barely integrated structure necessarily leads to heterogeneity of information on various levels. Different data formats and coding technics, the inconsistent use of data bases and intersections of redundant or differently defined functionalities of the single IT-systems represent typical characteristics.

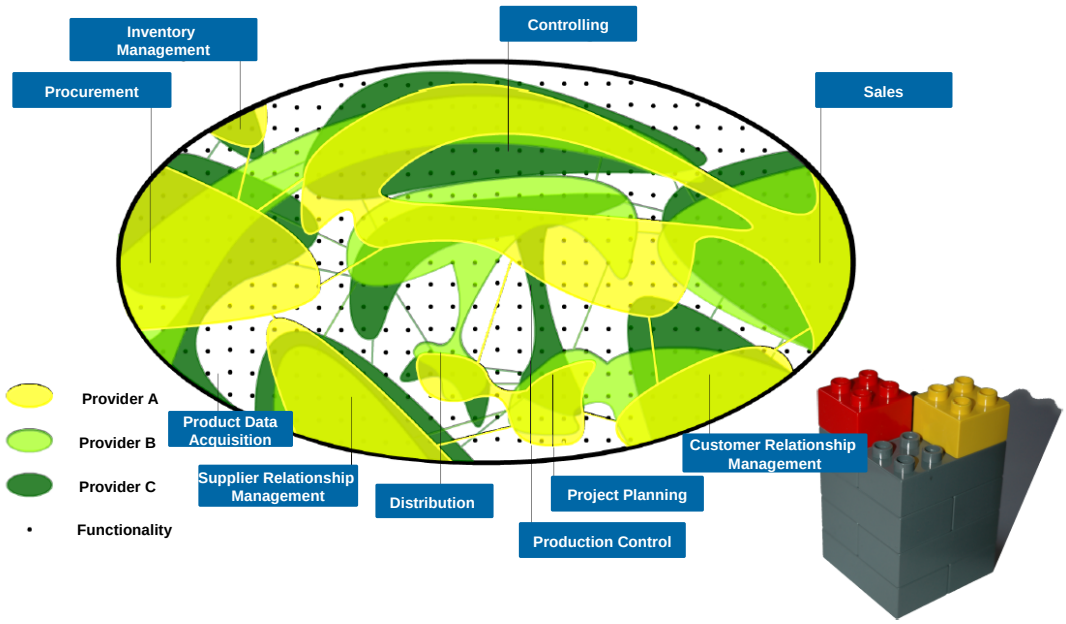


Figure 1: Differences in function-related design of ERP-system-landscapes

This makes it difficult if not impossible to detect and collect all information regarding a certain topic, may it be a certain product or order, throughout the IT-landscape and to prepare, illustrate and further process this information in a unique way. These deficits result in divergences between planning results and the company´s actual situation [4], [8], [9]. High stocks, heavily varying process times as well as frequent changes in order sequences illustrate these shortcomings [10], [11], [12], [13]. Since in many cases the introduction and adaption of IT-Systems is significantly more time- and cost-consuming than a reorganization of process structures, the IT is increasingly forming a serious "bottleneck" (cf. Figure 2). It narrows down the required flexibility of companies to quickly react to deviations or changes and to cope with the growing market dynamics.

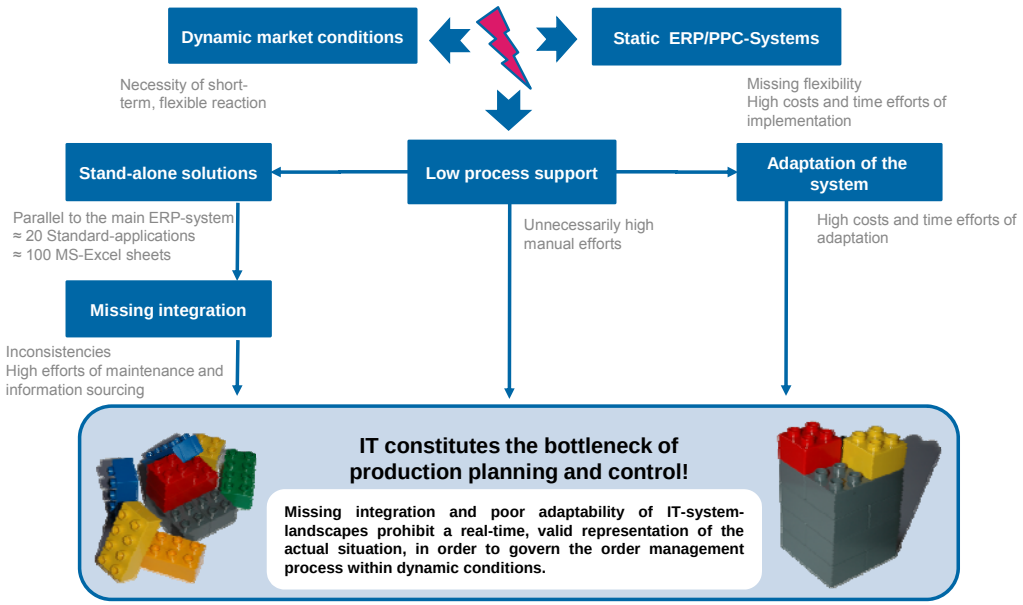


Figure 2: Conflict between dynamic market conditions and static IT-systems

2 Scientific Context

In recent years, a multitude of scientific disciplines deal with the design of production systems in order to increase the agility of companies [14], [15], [16]. Especially the transfer of control theory and cybernetics to production planning and control has gained a lot of attention with respect to operational production management [16], [17].

The idea can be traced back to an automated and continuous feedback-loop from running processes in order to establish a flexible production control with low disturbance sensitiveness [18]. A fast and appropriate reaction to deviations from the planned situation shall be enabled and therefore a better capability to deal with dynamics compared to classical production planning and control approaches [18]. Therefore information transparency plays a key role in today´s planning and controlling of a producing company.

However this transparency regarding the actual situation in a company is not achieved in practice yet. The rout causes lay in the described IT-structure of companies which doesn't allow for the integrated transfer of information from the point of creation to the point of control and its low adaptability [4], [9], [12], [18]. It can therefore be regarded as a basis for the realization of an agile enterprise to incorporate the perspective of information flows within the ERP-system in the approaches of the organizational design of agile processes and structures within a company and use this basis for the development of new approaches for an adaptive ERP-system-design. However this integrated perspective is still widely neglected.

3 Solution approach

As mentioned the increasing speed of information generation of today's business environment and the increasing need to quickly detect, acquire, store, process and illustrate information in an appropriate and consistent way requires new approaches for ERP-system design on the one hand. On the other hand this approach has to go hand in hand or rather has to be based on the design of the entire information system within companies considering all business objects that information is related to, like tasks, people, processes and IT-Systems, in a comprehensive manner.

Regarding the technical base for such a new paradigm of ERP-system design, the change of the IT-landscape in recent years becomes relevant. The client/server-model is being challenged by the service oriented architecture (SOA) – approach. SOA in general means the design of software systems build up from loosely coupled modules (services) with clearly defined tasks [19]. The single services package data and application logic and interchange messages over technically standardized interfaces [20], [21]. Through the configuration of highly standardized services it gets possible to design an individual overall system through workflows that are positioned along the business processes and that can easily be reconfigured and adapted. Differences in requirements are served by different service-configurations. That means that even though the single services are highly standardized the system configuration itself can be highly individualized.

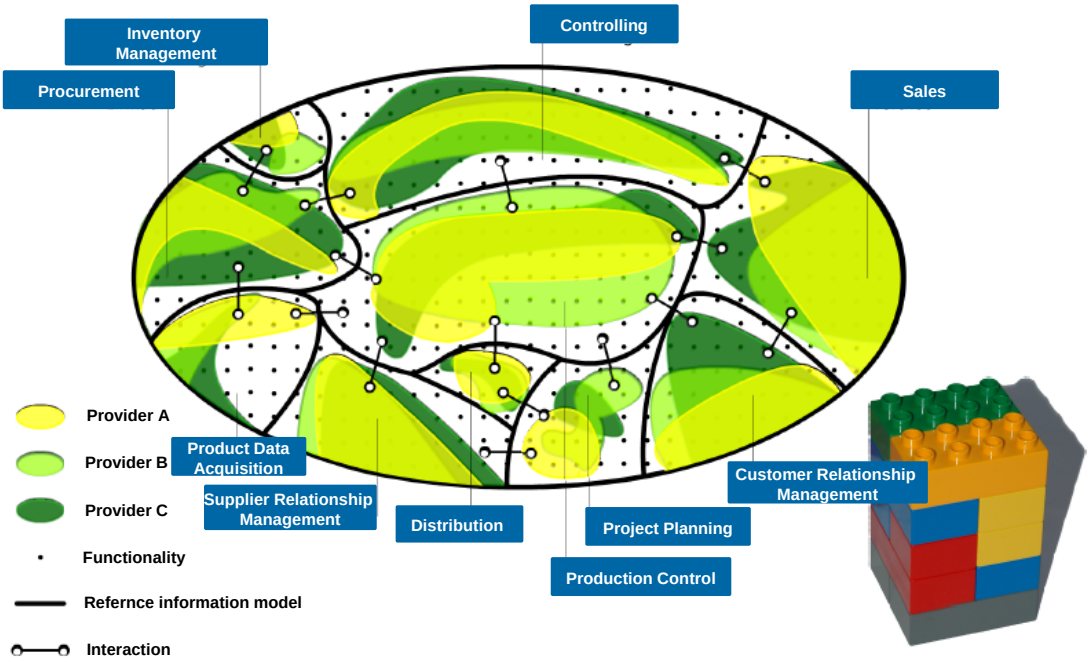


Figure 3: Reference model for the modularization of IT-system-landscapes

The applicability of this approach regarding the provision of ERP-related services and their configuration to a whole system is under wide discussion [19]. It would mean the split of the monolithic ERP-block to service-modules which can be more easily (re-)configured according to companies' needs compared to classic software systems.High potential can be seen in an IT-provider-spanning compatibility of the modules which would allow for a best of breed competition on the provider side and an according solution on the company-side. The software providers would gain a much broader market access and economies of scale since their modules would also be applicable for companies which nowadays are restricted to one ERP-system of one provider, although parts of that system may not fit their requirements. Accordingly companies could select modules and configure their individual system.
However this requires set rules and standardized data-related interfaces regarding the design and interrelation of these modules (cf. Figure 3). It doesn't seem appropriate or even manageable for companies to orchestrate their own ERP-System based on single functional services. This approach will only be applicable in practice if modules include a certain breadth of functionality keeping the number of different modules manageable and thus make a configuration of modules for companies possible. While the technical compatibility is already given within the framework of e.g. an enterprise service bus linking the different modules to each other via the routing of messages and function calls, two main aspects have to be dealt with within this research approach (cf. Figure 4):

- Which functionalities are clustered in a certain module? What is the frame of a certain module?
- What content in form of data is exchanged in order for the module to be able to perform the intended functionalities and what data does it offer to other modules?

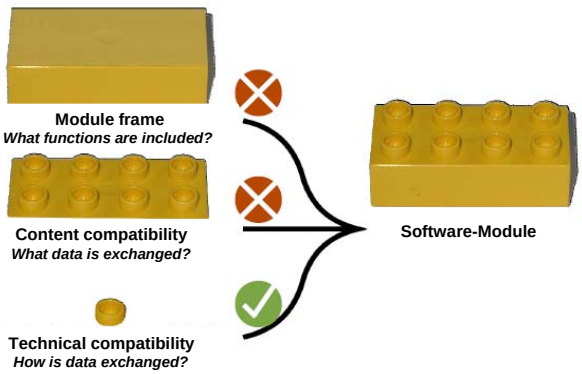


Figure 4: Module composition within service oriented architectures

Accordingly the research design is divided in 2 parts. Part one focuses on the data and information interchange within the order management process in producing companies.

As the situation analysis clearly characterized the main problem in the field of information integration this means, that the solution approach has to be built upon the network of information flows within companies. The composition of an information system of a company requires a precise concept of the objects and circumstances that the information is related to. The tasks of information management have to be linked to the IT-supported part of the information system as well as to the underlying physical and administrative processes in order to design a consistent overall system.

Therefore a comprehensive information model is being set up that relates process activities to their supporting IT-functionalities and corresponding data in- and outputs. The model is based on the Aachener PPC-model, a widely accepted model for production planning and control in the engineer-to-order industry [Quelle].

Within this framework different ontologies for the single objects of the inner-company information system will be set up, namely a data-ontology, an IT-function ontology, a task ontology and an ontology for business rules. These ontologies serve as semantically unified bases for a shared consistent understanding of the single elements and their context within the information system.

These ontologies will be linked to the Aachener-PPC process model using an approach from Thomas & Fellmann [22] for an ontology based business process modeling. The result is a reference information model aligned along the process activities of order management from the offer preparation to distribution of the final product. For each process step the linked ontological elements give an exact description of the input-data, the required IT functionality, the business rules which are employed and the task within production planning and control that is fulfilled. The semantic annotation and precision of the ontological design is necessary to prevent room for interpretation of the meaning of the single elements and to create a shared understanding of the context which is meant to be the later basis for service-module development.

Accordingly, in a second step this information model is used for the application of cluster-methods that analyze the interrelations of data flows and calculate functional correlations based on mathematical models in order to set validated frames for the grouping of functionalities in single modules.

The resulting reference-model consists of comprehensibly compatible, process-related ERP-modules with open, standard interfaces of information transfer. The context of

each element within the information model can be ontologically traced within the respective context and guaranties a unified understanding of the interrelations. This can be seen as a prerequisite for the interoperability of software modules which are developed by different groups of programmers. If software provider apply this model in the design of their service-modules, it will enable companies to be more flexible in designing and adapting their IT-structures to changing business conditions and therefore allow for a more efficient and competitive performance throughout the industry.

4 References

[1] Lau, C. (2010): Methodik für eine selbstoptimierende Produktionssteuerung. Utz, München, 2010.
[2] Westkämper, E., Zahn, E. (2009):"Wandlungsfähige Produktionsunternehmen: Das Stuttgarter Unternehmensmodell. Springer, Berlin 2009.
[3] Nyhuis P. (2008): Beiträge zu einer Theorie der Logistik. 1. Aufl., Springer, Berlin 2008.
[4] Wildemann, H. (2008): Entwicklungslinien der Logistik. In: Nyhuis, P.: Beiträge zur Theorie der Logistik. Springer: Berlin 2008, pp. 19-41.
[5] Schuh, G., Westkämper, E. (2006): Liefertreue im Maschinen- und Anlagenbau: Stand, Potenziale, Trends. Stuttgart 2006.
[6] Steven, M. (2005): Supply Chain Management für globale Wertschöpfungsprozesse. Wirtschaftswissenschaftliches Studium, Beck, München 2005, pp. 195-200.
[7] Wiendahl, H.(2006): Erfolgreiches Produktionsmanagement im Mittelstand. Shaker, Aachen 2006.
[8] Windt, K. (2006): Selbststeuerung intelligenter Objekte in der Logistik. Selbstorganisation, Böhlau, Köln 2006, pp. 271-314.
[9] Fleisch, E., Christ, O., Dierkes, M. (2005): Die betriebswirtschaftliche Vision des Internets der Dinge. In: Fleisch, E., Mattern, F. (2005): Das Internet der Dinge. Ubiquitous Computing und RFID in der Praxis:Visionen, Technologien, Anwendungen, Handlungsanleitungen.Springer, Berlin 2005, pp.3-37.
[10] Wiendahl, H.-H. (2006): Erfolgreiches Produktionsmanagement im Mittelstand. Shaker, Aachen 2006.
[11] Wildemann, H. (2006): Lean als Paradigma produzierender Unternehmen. 3. Lean Management Summit: Aachener Management Tage 2006. Aachen 2006, pp. 133-141.
[12] Windt, K. (2006): Selbststeuerung intelligenter Objekte in der Logistik. Böhlau,Köln 2006,pp. 271-314.
[13] Scholz-Reiter, B., De Beer, C., Freitag, M., Rekersbrink, H., Tervo, J. (2008): Dynamik logistischer Systeme. In: Nyhuis, P. (2008): Beiträge zu einer Theorie der Logistik. Springer, Berlin 2008, pp. 109-138.
[14] Wiendahl, H.-H., EL Maraghy H., Nyhuis P., Zäh M., Duffie N., Brieke M. (2007): Changeable manufacturing - classification, design and operation. In: CIRP Annuals. Manufacturing Technology 56 (2): S. 783–809.
[15] Dyckhoff, H. (2003): Grundzüge der Produktionswirtschaft: Einführung in die Theorie betrieblicher Wertschöpfung. Springer, Berlin 2003.
[16] Brosze, T. (2011): Kybernetisches Management wandlungsfähiger Produktionssysteme. Dissertation RWTH Aachen. Apprimus, Aachen 2011.
[17] Meyer, M. (2007): Logistisches Störungsmanagement in kundenverbrauchsorientierten Wertschöpfungsketten. Shaker, Aachen 2007.
[18] Gierth, A. (2006): Beurteilung der Selbststeuerung logistischer Prozesse in der Werkstattfertigung. Dissertation RWTH Aachen. Shaker, Aachen 2006.
[19]Mertens P. (2005): Grundzüge der Wirtschaftsinformatik. 9. Aufl., Springer, Berlin 2005.
[20] Marks, E. A., Bell, M. (2006): Service-oriented architecture. A planning and implementation guide for business and technology. Wiley, Hoboken, N.J 2006.
[21] Woods, D., Mattern, T. (2006): Enterprise SOA. Designing IT for business innovation. O'Reilly, Sebastopol, CA 2006.
[22] Thomas, O., Fellmann, M. (2007): Semantic EPC: Enhancing Process Modeling Using Ontology Languages. In: M. Hepp, K. Hinkelmann, D. Karagiannis, R. Klein und N. Stojanovic (Hg.): Semantic Business Process and Product Lifecycle Management. Proceedings of the Workshop SBPM 2007. SBPM. Innsbruck, 7. April 2007 (CEUR Workshop Proceedings).

Threshold model of diffusion: An agent based simulation and a social network approach

Suk-ho Kang, Wonchang Hur. Jeehong Kim, and Daeyoung Kim

***Abstract*— Innovation diffusion is a social process in which an innovation is adopted by the members of a social system. But, why some diffusion processes do precipitate bandwagon dynamics and others fail to do? In addressing this question, we considered a threshold model of diffusion. We particularly focus on the patterns of threshold arrangement: how individuals are arrayed over a given network topology. Employing a multi-agent simulation based on a threshold model of diffusion, we found that the model always results in the one of the two following states irrespective of the randomly varying diffusion networks; almost all the adopters will participate in adoption or only the small part of them will do so. In addition, we proposed 10 measures capturing the patterns in which individual threshold levels are associated with its topological features in the given network. From the regression analysis, it turns out that the nodes located closer to the innovators play a critical role in promoting diffusion.**

***Keywords*—Agent based simulation, diffusion, social network, threshold model, centrality**

I. Introduction

INNOVATION diffusion is a social process in which an innovation is adopted by the members of a social system. Diffusion studies have provided many substantive examples showing that most successful innovations have an S-shaped rate of adoption [1]. The increasing rate of diffusion in the early period and the subsequent decreasing rate in the later period are characterized as snowball effect or chain reaction [2]. The chain reaction indicates the effect of interpersonal communication on the process of diffusion, the nature of which is a positive feedback loop: increases in the number of adopters create stronger pressures, and stronger pressures, in turn, cause increases in the number of adopters [3].

An interesting concept regarding this chain reaction of diffusion process is the critical mass. The notion of critical mass originated in physics, where it was defined as the amount of radioactive material necessary to produce a nuclear reaction [4], [5]. In the context of diffusion studies, the critical mass occurs at the point at which enough individuals in a system have adopted an innovation so that the innovation's further rate of adoption becomes self-sustaining [1]. Due to the chain reaction, there is, after the small number of system members adopts an innovation, relatively rapid adoption by the remaining members and then a period in which the holdouts finally adopt [2].

This paper raises the following questions on these features of diffusion dynamics; what are the key factors to instigate the positive feedback loop that makes diffusion self-sustaining? And how they are related to the other part of the system? Although these questions are fundamental to understanding why some diffusion processes do precipitate bandwagon dynamics and others fail to do, they are not largely explored in the diffusion literature.

This paper employs a multi-agent simulation based on a threshold model of diffusion. Threshold models are important in diffusion studies because it can easily describe a complex interdependency of individual decisions over time. Based on the threshold hypothesis, we consider a large number of networks generated random rewiring of individuals of different characteristic. We particularly focus on the patterns of threshold arrangement: how individuals are arrayed over a given network topology. This is different from the topologies of network, which is only related to the shape of a network. Then we investigated under what conditions diffusion processes are able to reach the critical mass, that is, becomes self-sustaining. Our study reveals that the intricate combination of threshold heterogeneity and interdependency is a crucial factor affecting the bandwagon dynamics. The results carry interesting implications on the way different peoples are connected with each other and its effect on diffusion dynamics.

II. The Related Literature

An important diffusion model that can easily describe a diffusion mechanic like chain reaction is a threshold model. Threshold models assume that individuals take account of how many others are behaving in a particular way in making their own decisions about participating in the behavior. The crucial part of this assumption is that they have different thresholds – that is, some individuals will adopt after only a small proportion of their alters has adopted, while others will not adopt until a large proportion of their alters has adopted [6]. As a result, the

Suk-ho Kang is with the Department of Industrial Engineering, Seoul National University, 599 Gwanak-ro, Gwanak-gu, Seoul 151-742, Republic of Korea (corresponding author. phone: 822-880-7173; fax: 822-889-8564; e-mail: shkang@snu.ac.kr).

Wonchang Hur is with College of Business Administration, Inha University 253 Yonghyun-dong, Nam-gu, Incheon 402-751, Republic of Korea (e-mail: wchur@inha.ac.kr).

Jeehong Kim is with the Department of Industrial Engineering, Seoul National University, 599 Gwanak-ro, Gwanak-gu, Seoul 151-742, Republic of Korea (e-mail: jivong97@snu.ac.kr).

Daeyoung Kim is with the Department of Industrial Engineering, Seoul National University, 599 Gwanak-ro, Gwanak-gu, Seoul 151-742, Republic of Korea (e-mail: kdy555@snu.ac.kr).

distribution of thresholds in a population becomes a key factor closely linked to the mechanic of diffusion process and comes to affect significantly the final extent of the diffusion process. In fact, Reference [7] demonstrated mathematically that a slight perturbation of the threshold distribution sometimes generates entirely different level of participation. In this way, threshold models provide an important explanatory mechanism on why diffusion causes a chain reaction to start and various proportions of a collectivity's members to adopt.

The threshold concept has been adapted, extended, and incorporated frequently into many later diffusion models as an important theoretical background for them. For example, [4] points out the problem of identifying individual thresholds which requires detailed knowledge of the individual expected rate of return on investments in public goods. In order to address this problem, the authors introduce a stochastic learning model in which individual thresholds are not deterministic. In the model, the structural configuration of thresholds is treated in two stylized cases, strong ties and weak ties, and their simulation results support the strength-of-weak-ties hypothesis. Reference [8] tries to relax the assumption of non-reciprocal communication of classic threshold models, and points out that a person learns about his/her neighbors' preferences, their willingness to participate, and does not directly respond to their actions. The author argues that people with low thresholds, who are highly predisposed toward participation, are affected much more by social position than people with high thresholds. The author also shows that how strong links can be better for participation when thresholds are low and weak links can be better when thresholds are high. The model is well aligned with our intuition that whether a low-threshold person participates or not depends greatly on whether that person happens to have some sympathetic friends, while a high-threshold person participates only if a great mass of people participate. Reference [3], by employing a "core-periphery" network, highlights the role of boundary agents. The study introduces the two important concepts, boundary pressure point (agents with high threshold and high connectivity) and boundary weakness (agents with low threshold and low connectivity), and proposes that they have relatively greater effect on diffusion extent, particularly in lower-density networks.

Most of these models relax the original model's assumption that an actor's decision to join a social movement has the same effect on all other actors in the social system. This relaxation of the restrictive assumption allows them to analyze the structural effect, the effect of preexisting social networks linking the actors in which the flow of influence is restricted and channeled. The difference between the models largely reflects differences in the theoretical understanding of how the structural arrangement of social ties into a network affects social influence.

One recent work by [9] advances this issue further into the more sophisticated level. In contrast to previous work on the structural configuration of networks, the author focuses on how thresholds are arrayed across a network and shows that it can profoundly affect diffusion dynamics. The author, from abstract computational experiments, finds that a balance of similar and dissimilar thresholds is important in maximizing participation. That is, the optimal distribution of thresholds across networks would be a pattern where agents associate with a certain proportion of others with similar threshold values while keeping a certain level of friendship with others of discrepant thresholds.

Our work starts from the same motivation of his work, but differs significantly from it in several points. First, we assume a normal distribution of threshold distributions rather than the quasi-uniform distribution adopted by many previous models. Although there has been little evidence on the exact distribution of threshold, it is generally accepted that thresholds are positively associated with the time-of-adoption which is found to be follow a bell-shaped distribution. Second, we consider scale-free network, which has more empirical supports as a stylized model for many real world social networks, but has not been treated much in the network threshold models. Most important, our work takes a more exhaustive approach than the previous works. We consider all the possible threshold arrangements over given network topologies rather than generate networks with intended arrangement patterns. We undertake statistical analysis for these exhaustive enumerations of arrangement patterns and seek to find any patterns that may be related with the diffusion dynamics.

III. The Model

Consider a process in which a certain product diffuses across a social network of potential adopters. Each adopter in the network makes a decision of whether to adopt the product based on a simple condition; whether the proportion of adopters among his (or her) network peers at time *t* is above his or her *threshold*. Previous diffusion research has conceptualized the proportion of adopters among network peers as *exposure*, which represents a level of peer pressure imposed on a potential adopter. Thus the adoption condition implies that adoption occurs when peer pressure on an adopter exceeds threshold. The decision rule is described in Table 1. In the table, n_i is the number of neighbors in the network, θ_i the threshold of agent *i*, and $n_{i\_a\_A}(t)$ the number of adjacent adopters that have adopted the product A at time *t*.

Unlike the original model, we assume that influence flow among agents is restricted by the structure of a preexisting social network. That is, each agent adopts the product when the proportion of agents who have already adopted among network peers is above his/her threshold. The agents' threshold values are assumed to be determined from a normal distribution. We assume that the distribution parameters, μ and σ, are not independent. This is intended to have the fixed proportion of

TABLE I
Decision Rule

Adoption criteria	Adoption result
$n_{i\_a\_A}(t) \geq n_{i\_*}\,\theta_i$	Adopt A
$n_{i\_a\_A}(t) < n_{i\_*}\,\theta_i$	Don't adopt

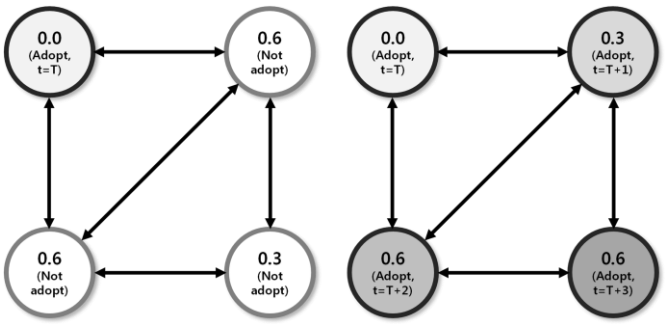


Fig. 1. Two same networks with different threshold arrangement

0-threhold agents in the population regardless of the threshold distribution. We consider 2.5% of the population as the initial adopters. This percentage is borrowed from the adopter categorization by [1], in which he reportedly says that the category that adopts at the earliest time makes up about 2.5% of the population. That is, diffusion starts from the 2.5% of the population (initiators, hereafter) and stops when there are no additional adopters. The model output, the diffusion extent, is thus the number of adopter when diffusion stops.

The diffusion extent is determined from 3 variables; threshold distribution, network topology, and threshold arrangement. Threshold arrangement refers to a way in which agents are arrayed over a given network. It is important to understand that even when threshold distribution and network topology is fixed, the arrangement patterns can be varied substantially and the model produces a completely different outcome depending on the pattern. The following example demonstrates the effect of the arrangement pattern. In Fig. 1, there are two different networks of 4 agents; their thresholds are 0%, 30%, 30%, and 60%, respectively. Although the two networks are of the exactly same topological type, we can see that their diffusion outcomes are completely different – diffusion extent of the left is 1, but the right is 4.

Notice that although threshold re-arrangement (hereafter rewiring) does not change the network-level topological features, it does changes the micro-structure of how individuals interact with each other. That is, although rewiring does not change degree distribution, network centralization, clustering coefficients, or many other network-level features, it does change individual neighbors' profile in terms of their threshold values. That is, the difference in diffusion extent will be caused by the difference in the relative position of adopters in the given network. This means that the factors affecting diffusion should be able to capture the characteristics in how adopters are distributed and positioned in a given network topology and the patterns that individual thresholds are correlated with the topological features.

In order to examine the characteristics of the diffusion process, an agent-based model that simulates the specified adoption behavior has been built. For simulation, we first create a population of 1,000 potential adopters and their threshold values were assigned from a normal distribution. We vary threshold distribution by varying its μ from 0.25 to 0.4. For network topology, we considered 3 random networks and 3 scale-free networks with varying average degrees. For each network topology, simulation was conducted on 1,000 different threshold arrangements generated by rewiring the adopters randomly under a given network topological type. Note that only one outcome is obtained from one network setting since the adoption process is deterministic.

Fig. 2 shows the effect of threshold distributions on the diffusion extent. The y axis denotes the average diffusion extents obtained from the different arrangement. Expectedly, as μ increases, the average of diffusion extents tends to decrease. When μ is large enough (e.g. $\mu > 0.4$ in case of random networks), the diffusion extents become always 0%. Similarly, when μ is small enough, diffusion reaches 100% regardless of the thresholds arrangement. That is, there is a range of μ in which the diffusion extents can vary depending on how thresholds are arrayed over the networks.

Then, we examine how the diffusion extent varies depending on the threshold arrangements when μ is given in the range. Fig. 3 shows the distribution of the diffusion extents resulted from simulation on 1,000 diffusion networks when μ is 0.35 and σ is 0.18. From the previous graphs, we see that when population thresholds follow a normal distribution with μ=0.35 and σ=0.18, the average extents are about 20% ~ 40% depending on the network types. But Fig. 3 shows that its distribution is clearly bifurcated. That is, even though the population threshold has a fixed distribution, diffusion will result in one of the two extreme cases depending on how threshold are arranged over the given network; almost all the adopters (more than 95%) will participate in adoption or only the small part of them (less than 30%) will do so. It is surprising that there are no such cases that 50% ~ 60% of individuals adopt the behavior at the end of simulation although diffusion networks were rewired randomly.

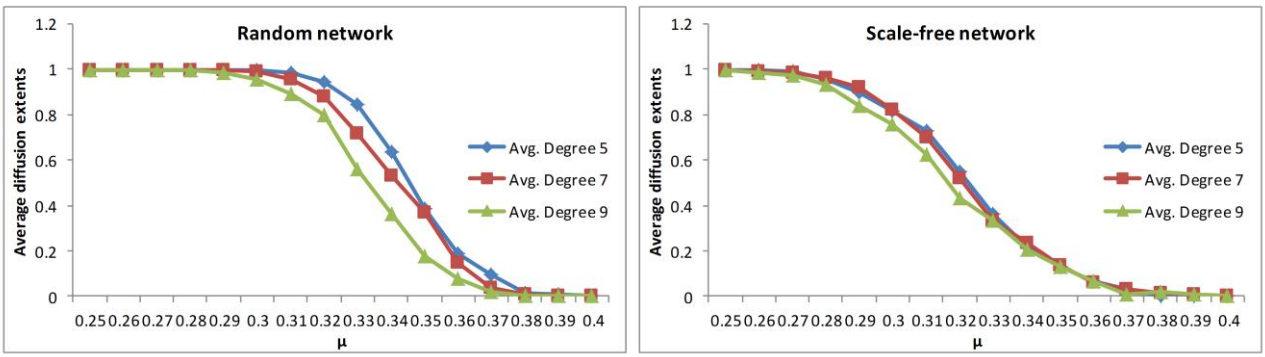


Fig. 2. Diffusion extent according to threshold distributions

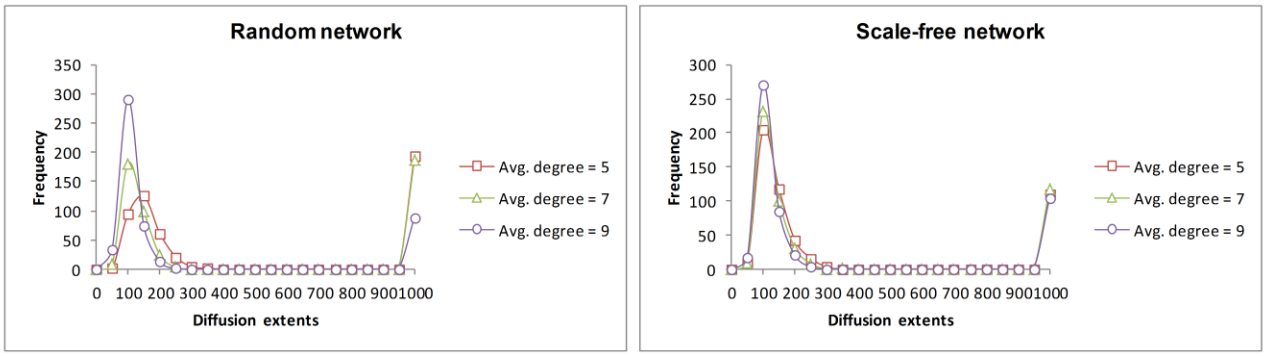


Fig. 3. Distribution of diffusion extents ($\theta \sim N(0.35, 0.18)$)

The finding raises two significant questions. First, why does the diffusion extent always fall into one of the two extreme ranges, that is, above 95% or below 30%? It is certain from the result that once a small portion of adopters (about 20~30% in this setting) participates in the behavior, then all the remaining members will eventually follow them. This finding is consistent with the well-known theory of the critical mass from the literature on social dynamics. The theory postulates that collective actions characterized by population heterogeneity and interdependence, which are well captured in our behavior model, are often activated by the small portion of early movers. From the perspective of the critical mass theory, we can consider the 10~20% early adopters as the critical mass of participants that has to be crossed in order for perfect adoption to occur. Our simulation proves that the small part of the population does lead to the unanimous action, and shows how such behavioral dynamics can be derived from the simple behavior rules.

The second question, which this paper lays more focus on, is what factors contribute to perfect diffusion? Regarding this question, it is important to understand that a network's initial topology remains unchanged during the rewiring process. That is, 1,000 diffusion networks, although their diffusion outcomes are completely bifurcated, are almost identical in terms of their topological features. For instance, when a network is given, we can create another network by simply relocating nodes in the given network. Notice that the new network generated this way is isomorphic to the original one, meaning that these two networks are exactly same in every topological feature. However diffusion outcomes on these two networks can be significantly different because the way nodes interact with each other is different in each network despite their topological isomorphism.

IV. Factors Contributing to Perfect Diffusion

So far we have described the diffusion model and investigated its behavior. Although the proposed model is based on a quite simple behavior rule, it generates the results that cannot be easily explained. Why does the diffusion extents are bifurcated depending on the pattern of threshold arrangement? This result suggests that there are a certain proportion of agents that should be crossed over in order for diffusion to reach the whole population. In other words, once diffusion reaches the proportion, it then continues like a self-sustaining chain reaction process leading to the almost perfect adoption regardless of how the remaining agents are arranged. Therefore we need to explain when diffusion can establish the necessary initial contributors. The simulation results suggest that there are particular ways of arranging agents according to their thresholds, which are advantages for establishing the initial adopters sufficiently for precipitating the self-sustaining chain-reaction.

There are some important facts that must be related with the role of threshold arrangement in diffusion. First, diffusion starts from the 2.5% initiators so the initial adopters should be located around them. Second, low-threshold nodes are likely to adopt earlier than others so it is advantageous for them to be placed near the initiators. Third, some nodes are topologically critical in promoting diffusion. From these facts, it is supposed that following factors are likely to be associated with promoting diffusion.

A. Position of the starting nodes

The simulation result shows how initial contributions can precipitate a chain reaction that may ultimately spread to every member of the group [4]. Since diffusion starts from the initiators, where and how they are located in a network should have a significant effect on the diffusion outcome. We assume that it would enhance diffusion if initiators are evenly dispersed over a network to cover a wider area in a cooperative way

B. Threshold values of nodes located close to the starting nodes

Since diffusion proceeds along the path from initiators, nodes closer to them are considered as candidates for prospective adopters earlier than others. Since exposure must be low in the early time, if they have high thresholds it is unlikely that they become adopters. This should blockade the path from initiators to other nodes in a network and then the overall diffusion process is likely to be deterred. In this respect, we can hypothesize that it would be advantageous that agents with low threshold are located near the initiators, and, similarly, agents with higher thresholds are located farther from them. If a diffusion network is structured in that way, then adopters' thresholds becomes monotonically increasing from the nearest initiators to other nodes.

TABLE 2
MEASURES FOR A DIFFUSION NETWORK

Factors	Measures	Meaning
Thresholds according to a node's distance to the starting node	M1. Average thresholds of nodes adjacent to the initiators M2. Average thresholds of nodes at 2-edges apart from the initiators M3. Average thresholds of nodes at 3-edges apart from the initiators	Whether early adopters are located close to the initiators?
Position of the initiators	M4. Average degrees of the initiators M5. Average closeness of the initiators M6. Average betweenness of the initiators M7. Average distances between the initiators	The coverage and distribution of the starting nodes
Thresholds according to a node's centrality	M8. Correlation between degree and threshold M9. Correlation between closeness and threshold M10. Correlation between betweenness and threshold	Whether central positions play an expected role of promoting diffusion?

C. Threshold values of nodes with high centrality

In addition to nodes closer to initiators, central nodes must be also important for promoting diffusion. Central nodes are those that are extensively involved in relationship with others. These nodes have high level of access to others and play an important role of brokering information from one part of a network to the other. Hence, it is important that highly central nodes have low thresholds so that they adopt early and help adoption expand to a wider area.

Based on the discussion so far, we consider the following measures for the abovementioned factors. Table 2 provides a brief description of those measures with the equations to calculate them.

V. RESULTS

Employing factors as independent variables, we performed a Probit regression to examine their effects on diffusion. As a dependent variable, we employ a binary variable P indicating whether the diffusion extent was over 90% or not. Table 3 shows the regression results, which clearly indicate that when the neighbors of initiators (M1) and these agents' neighbors (M2) have low thresholds, the extent of diffusion can be significantly enhanced. The tables show that only M1 and M2 were consistently significant regardless of k and μ. This implies that the most significant factor affecting diffusion extents is the average threshold of the agents who are incident to or located closer to the initiators. Clustering of low-threshold agents around the initiators facilitates the formation of groups of early adopters, who spread their influence across the network. This implication pertains to both random networks and scale-free networks.

Other variables, by contrast, turned out to be mostly insignificant. Two exceptions are M4 and M8 in some scale-free networks. It is known that in scale-free networks, the degree distribution follows a power law, which means that the vast majority of nodes are those with small degrees, with only a few having relatively high degrees. The highest-degree nodes are often called "hubs", and are thought to serve specific purposes in diffusion. Our results imply that whether hubs have low thresholds or not is a significant factor affecting diffusion extension.

VI. CONCLUSION

In this paper we considered a generic threshold model of diffusion and explored how it behaviors under the various diffusion networks by using agent-based simulation. Despite its simplicity, the model has some interesting characteristics worth exploring. First, adopters are heterogeneous in that they have their own varying threshold levels. As a result, each adopter behaves differently even under the same level of exposure. Second, more importantly, adopters behave based on local information that comes only from their network partners; the adoption behavior of their network partners in a population. This implies that diffusion will be largely affected by the network structure of who is connected to whom in the given population.

TABLE 3
PROBIT ANALYSIS OF DIFFUSION NETWORK

Random Network

Degree	5		7		9	
μ	0.35		0.35		0.34	
Variable	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
Intercept	-50.38	22.08	-24.75	23.54	-55.07	25.23
M1	25.75***	4.62	34.88***	5.31	38.68***	5.86
M2	51.76***	8.31	66.25***	10.36	60.18***	9.90
M3	30.21***	7.94	25.61***	6.99	2.61	4.30
M4	0.03	0.06	-0.04	0.07	0.20***	0.07
M5	-0.74	0.60	0.22	0.62	1.36**	0.64
M6	44.61	72.29	-62.21	73.98	-7.58	78.39
M7	-60.05	380.42	-728.73	618.86	-2521.12***	899.89
M8	20.06**	9.22	12.70	9.86	7.85	11.39
M9	4.43	6.84	-4.74	5.37	-2.55	5.95
M10	-11.08	8.06	1.41	7.66	-6.76	8.44
Avg. extent	461.112		362.623		390.038	
R square	0.2746		0.2602		0.2037	

Scale-free network

Degree	5		7		9	
μ	0.33		0.33		0.33	
Variable	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
Intercept	-3.34	19.85	-28.54	21.07	-67.41	22.25
M1	37.76***	5.30	25.00***	5.44	31.96***	6.50
M2	37.78***	9.33	41.50***	10.66	42.06***	11.76
M3	27.17***	8.07	7.11	7.32	17.93***	6.70
M4	-0.08	0.06	0.03	0.07	0.14	0.07
M5	-0.33	0.30	-0.27	0.21	-0.59***	0.19
M6	-87.69	58.67	12.70	56.11	98.43	55.06
M7	20.39	181.35	-81.43	178.73	201.13	208.51
M8	25.43***	8.69	22.12**	8.94	11.41	9.45
M9	-5.61	3.93	-3.87	4.76	-1.36	5.68
M10	-9.43	6.95	-10.90	6.64	3.78	6.45
Avg. extent	431.822		397.364		393.484	
R square	0.2629		0.2602		0.3334	

Notes: ***Significant at p<0.01; **Significant at p<0.05; *Significant at p<0.10

These features of individual behavior, heterogeneity and locality, are generally believed to contribute to the unexpectedness of collective behavior considerably. Consistent with this expectation, the simulation results showed that the model always results in the one of the two following states irrespective of the randomly varying diffusion networks; almost all the adopters (more than 95%) will participate in adoption or only the small part of them (less than 30%) will do so. This result suggests the model's behaviors are consistent with the well-known hypotheses from the theory of the critical mass. In addition, we proposed 10 measures capturing the patterns in which individual threshold levels are associated with its topological features in the given network. From the regression analysis, it turns out that the nodes located closer to the innovators play a critical role in promoting diffusion.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No. 20110016160)

REFERENCES

[1] E. M. Rogers, "Diffusion of Innovations," 5th edition, New York: Free Press, 2003.
[2] A Nypan, "Diffusion of innovation and community leadership in east Africa," *Acta Sociologica*, Vol. 13, No. 4, pp. 253-268, 1970.
[3] E. Abrahamson and L. Rosenkopf, "Social network effects on the extent of innovation diffusion: A computer simulation," *Organization Science*, Vol. 8, No. 3, pp. 289-309, May 1997.
[4] M. W. Macy, "Chains of Cooperation: Threshold Effects in Collective Action," *American Sociological Review*, Vol. 56, No. 6, pp. 730-747, Dec 1991.
[5] P. Oliver, G. Marwell and R. Teixeira, "A Theory of the Critical Mass. I. Interdependence, Group Heterogeneity, and the Production of Collective Action," *American Journal of Sociology*, Vol. 91, No. 3, pp. 522-556, Nov 1985.
[6] D. Krackhardt, "Organizational viscosity and the diffusion of controversial innovations," *Journal of Mathematical Sociology*, Vol. 22, No. 2, pp. 177-199, 1997.

[7] M. Granovetter, "Threshold models of collective behavior," *The American Journal of Sociology*, Vol. 83, No. 6, pp. 1420-1443, May 1978.
[8] M. S-Y. Chwe, "Structure and Strategy in Collective Action," *American Journal of Sociology*, Vol. 105, No. 1. Jul 1999.
[9] Y-S. Chiang, , "Birds of moderately different feathers: Bandwagon dynamics and the threshold heterogeneity of network neighbors," *Journal of Mathematical Sociology*, Vol. 31, No. 1, pp. 47-69, 2007

Detecting Community Structure in Networks Based on Ant Colony Optimization

Bolun Chen[1] **Ling Chen**[1,2] **Yixin Chen**[3]
[1]Department of Computer Science, Yangzhou University , Yangzhou, 225127, China
[2]State Key Lab of Novel Software Tech, Nanjing University, Nanjing, 210093, China
[3]Department of Computer Science Washington University in St Louis, USA

Abstract—*Community structure is a nature characteristic of many real networks. The problem of detecting and characterizing the community structure has attracted considerable recent attention. In this paper, we present a new algorithm based on ant colony optimization to detect community structure in networks. In the algorithm, artificial ants are used to travel on a logical digraph to construct solutions of community detection. Each ants chooses its path according to the pheromone and heuristic information on each path. We define an associate degree and use it as the heuristic information. The quality of solution obtained by each ant is measured by its modularity. The algorithm is tested on synthetic and real world networks. The experimental results show that our algorithm can achieve higher quality results than other methods.*

Keywords-component; detecting community structure; ant colony optimization;

1 Introduction

The structure of complex networks has atrracted considerable attention in recent years[1-2]. Modularity or community structure is nature characteristic in many real networks such as social networks[3-4], biological networks[5-6], technological networks[8-9], and cooperative network relationships [10]. The community structure consists of groups of nodes such that nodes within a group are much more densely connected to each other than to the rest of the network. Figure 1 shows a network with three communities, where sets of vertices with identical color naturally form a community which are densely connected. Beacause communities are relatively independent of one another struturally, it is believed that each of them may coreespond to some fundamental functional unit. For example, one can consider functional categories of protein-protein interaction networks, biochemical pathways in metabolic networks, sets of Web pages on common topics on World Wide Web, social groups with simlar interests in social networks. Identifying and analyzing such communities from a large network provides a means for functional disection of the network and sheds light on its its organizational principles.

In the past decade, identification of community structure has attracted much attention in various scientific fields. Many methods have been proposed and applied for specific complex networks.

A huge variety of community detection methods are based on the centrality measure of modularity. In 2004, Newman [11] first proposed the concept of modularity and a fast algorithm for detecting the communities. In community detection， the nodes of a network are clustered into several groups to form communities. To cluster the nodes into groups, we have agglomerative and divisive hierarchical clustering methods. The agglomerative method starts from a state in which each vertex is the sole member of one of *n* communities, and then repeatedly joins communities together in pairs, choosing at each step the join that results in the greatest increase in modularity. The progress of the algorithm can be represented as a "dendrogram," a tree that shows the order of the joins. Cuts through this dendrogram at different levels give divisions, the algorithm selects the best cut by looking for the maximal value of modularity. Newman also presented a GN algorithm [14] which is divisive hierarchical clustering method. The algorithm repeatedly removes edges from the network to split it into communities. In each iteration, the algorithm deletes the edge with the largest betweenness in the network so as to obtain the maximum modularity value. The process of deleting edges ends until all the edges are removed. The partitioning result with the maximum modularity value is selected. Agarwal et al. [17] proposed two-level algorithms based on rounding mathematical programs. The algorithms round solutions to linear and vector programs. By comparing the solution quality to the fractional solution of the linear program, a bound on the available "room for improvement" can be obtained. The vector programming algorithm provides a similar bound for the best partition into two communities. Ruan et al. [18] propose a heuristic algorithm QCUT, which combines spectral graph partitioning and local search to optimize modularity. Using QCUT as an essential component, they also proposed a recursive algorithm HQCUT to solve the resolution limit problem. HQCUT can successfully detect communities at a much finer scale or with a high accuracy.

Spectrum analysis is another important approach in community detection. Donetti et al. [20] proposed an algorithm for the detection of communities in complex networks. The algorithm exploits spectral properties of the graph Laplacian matrix combined with hierarchical clustering techniques. A. Capocci et al. [21] also developed a spectral analyzing based algorithm to detect community structure in complex networks. The algorithm takes into account weights and link orientation. Since the method can efficiently detect clustered nodes in large networks even when these are not sharply partitioned, it

turns to be specially suitable for the analysis of social and information networks.

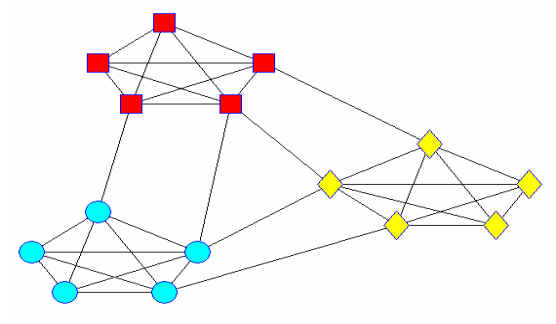


Figure 1 A network

Another effective approach for detecting communities in network is based on information theory. Rosvall et al. [22] developed an information-theoretic foundation for the concept of modularity in networks. They identify the modules of the network by finding an optimal compression of its topology, capitalizing on regularities in its structure. To comprehend the multipartite organization of large-scale biological and social systems. Rosvall et al. [23] also presented an information theoretic approach that reveals community structure in weighted and directed networks. They use the probability flow of random walks on a network as a proxy for information flows in the real system and decompose the network into modules by compressing a description of the probability flow.

Some community discovery algorithm are based on label propagation. Raghavan et al. [24] advanced a label propagation algorithm that uses the network structure alone as its guide and requires neither optimization of a predefined objective function nor prior information about the communities. In the algorithm every node is initialized with a unique label and at every step each node adopts the label that most of its neighbors currently have. In this iterative process densely connected groups of nodes form a consensus on a unique label to form communities.

In real world networks, some nodes naturally belong to several communities. Therefore, the study of overlapping community structures has attracted increasing attention recently, and many algorithms have been designed to detect such overlapping communities.

One way to detect overlapping communities is module penetration method. Palla1et al. [26] introduced an approach to analyzing the main statistical features of the interwoven sets of overlapping communities. Kumpula et al. [27] presented a sequential clique percolation algorithm (SCP) to do fast community detection in networks, for cliques of a chosen size. This method sequentially inserts the constituent links to the network and simultaneously keeps track of the emerging community structure. The SCP method allows for detecting *k*-clique communities at multiple weight thresholds in a single run, and can simultaneously produce a dendrogram representation of hierarchical community structure.

Fuzzy clustering is a useful approach for detecting overlapping community. Zhang et al. [28] presented an algorithm to identify overlapping communities in complex networks by mapping network nodes into Euclidean space and fuzzy c-means clustering. Ding et al. [29] also proposed an algorithm CDKFAP for detecting overlapping communities using a commute-time kernel based distance measure and fuzzy affinity propagation. By defining a new index measuring of the fuzziness of the nodes, the algorithm can rank and extract overlapping nodes of communities.

Another way to detect overlapping community is based on non-negative matrix factorization(NMF). Zhang et al. [30] presented a community detection method based on non-negative matrix factorization technique. Based on a popular modular function, a proper feature matrix from diffusion kernel is used in the algorithm. The algorithm can detect an appropriate number of fuzzy communities in which a node may belong to more than one community. The method can quantify how much a node belongs to a community. The quantification provides an absolute membership degree for each node to each community which can be employed to uncover fuzzy community structure. Using the non-negative matrix factorization method, Zarei et al. [31] also advanced an algorithm for detecting the structure of overlapping communities in complex networks. They introduced a vertex–vertex correlation matrix as the feature matrix of the NMF method.

In this paper, we present a new algorithm based on ant colony optimization (ACO) to detect community structure in networks. In the algorithm, artificial ants are used to travel on a logical digraph to construct solutions of community detection. Each ants chooses its path according to the pheromone and heuristic information on each path. We define an associate degree and use it as the heuristic information. The quality of solution obtained by each ant is measured by its modularity. The algorithm is tested on synthetic and real world networks. The experimental results show that our algorithm can achieve higher quality results than other methods.

2 Framework of the ACO algorithm for detecting community structure

A network can be presented as an undirected graph $G=(V,E)$, here V and E are respectively the sets of vertex and edges of graph G. Suppose G consists of n vertexes $v_1, v_2,\ldots,v_n$. Let $deg(v_i)$ be the degree of vertex v_i . In the ant colony optimization algorithm for detecting the community, each ant constructs a solution by searching directed by the pheromone and heuristic information. For the problem of community detection, a solution is represented by an n-dimensional vector $C=(c_1,c_2,\ldots,c_n)$, where $c_i=j$ indicates vertex v_i must be in the same community with its neighbor v_j.

For instance, a solution for the graph shown in Figure 2 be $C=(2,3,4,1,6,5)$ indicates there are two communities detected: $\{v_1,v_2, v_3,v_4\}$ and $\{v_5,v_6\}$. If a solution is $C=(2,3,4,1,4,5)$, then the two communities detected are $\{v_1,v_2, v_3,v_4\}$ and $\{ v_4,v_5,v_6\}$ which are overlapped since vertex v_4 belongs to two communities.

To construct a positive feedback, pheromone information τ_{ij} is assigned on each arc $E_{ij.}$ This pheromone information influences the choices the ants make: the larger amount of pheromone is on a particular path, the larger probability an ant selecting the path is. The intensity of pheromone information on the arc are updated in each iteration: it would be increased by the ants passing it and decreased by evaporation. Communications and cooperation between individual ants by pheromone information enable the ant colony algorithm to have strong capability of finding the best solutions.

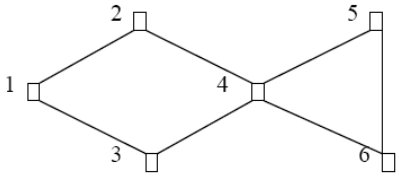


Figure 2 A graph

In the algorithm, the ants travel on a logical graph shown in Figure 3 to construct a solution. In Figure.3, the logical graph has n+1 vertexes V_1, V_2,…,V_{n+1} where V_1, V_2,…,V_n represents the vertexes v_1, v_2,…,v_n in the original graph and V_{n+1} is the nodes of termination. Starting from node v_1, each ant traverses the logical graph and its trace from V_1 to V_{n+1} forms a solution. Let the degree of v_i in the original graph be $deg(Vi)$, then there are $deg(Vi)$ arcs connecting two adjacent nodes v_i and v_{i+1}. Suppose there is an edge connecting V_i and V_j in the original graph, we add an arc E_{ij} connecting adjacent nodes v_i and v_{i+1} in the logical graph. If an ant at node V_i selects arc E_{ij} to reach V_{i+1}, then c_i=j in its solution C, which means V_i and V_j are assigned into the same community.

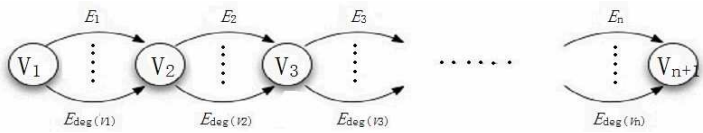


Figure 3 The logical digraph

The ant at node V_i select an arc E_{ij} to reach V_{i+1} according to a probability p_{ij}, which is determined by the pheromone information τ_{ij} and the heuristic information. It will take an ant n steps to complete a tour of traversing all the nodes. For every ant, its path traversing all the nodes forms a solution. Suppose there are m ants are used, m solutions can be obtained in each iteration. We evaluate their quality using the measurement of modularity. Pheromone on the paths included in the solutions with higher quality could have larger increment. The best solution with the highest quality found so far is denoted as $S_{best.}$ Pheromone on the paths included in the solution S_{best} will have the largest increment.

The framework of the algorithm is as follows.

Algorithm ACODCS (ACO for Detecting Community Structure)

Input: A: The adjacency matrix of logical digraph;

Output: The solution of the community structure detected;

Begin

1. Parameter initialization, set the initial value of parameters τ_{ij};
2. **While** not termination condition **do**
3. Starting from v_0, the m ants traverse on the logical graph according to the probability formula on each node. After all the m ants reach the node v_{n+1}, m solutions of community detecting are formed;
4. Evaluate the modularity of the m solutions;
5. Update the pheromone on each arc;
6. Select the solution with the highest fitness value found so far as S_{best} ;
7. **endwhile** ;
8. Output the result;

End

3 Details in implementation of the algorithm

In this section, we illustrate the technical details in implementing the algorithm ACODCS.

3.1 Probability for ants' path selection

On line 3 of the algorithm, ants at node V_i selects an arc to reach V_{i+1} according to a probability. At node V_i, there are $deg(V_i)$ arcs linking next node V_{i+1}, we use p_{ij}^k to denote the probability for ant k on node V_i to choose the path E_{ij}:

$$p_{ij}^k = \frac{\tau_{ij}^{\alpha} \cdot \eta_{ij}^{\beta}}{\sum_{j=1}^{\deg(V_i)} \tau_{ij}^{\alpha} \cdot \eta_{ij}^{\beta}} \quad (1)$$

Here, τ_{ij} is the pheromone on the arc E_{ij} between nodes V_i and V_{i+1}, it reflects the potential tend for V_i and V_j being in the same community. η_{ij} is a heuristic function which is defined as the visibility of the arc E_{ij}. We use the associate degree of the edge linking V_i and V_j in the original graph as the heuristic function of E_{ij} . Parameters α, β determine the relative influence of the trail information and the visibility.

3.2 The associate degree between the nodes

In (3) η_{ij} is a heuristic function which reflects the potential tendency for the ants to select the arc E_{ij}. In the graph of the network, if two nodes have large number of other nodes they both directly or indirectly connect with, then they have higher chance to be in the same community. Therefore we use the connectivity of two nodes V_i and V_j as the heuristic function η_{ij} on arc E_{ij}. The connectivity of two nodes is related to the surrounding nodes they connected with. Let the adjacent

matrix of the graph be $A=[a_{ij}]$, and $A^k=[a_{ij}^{(k)}]$. Obviously $a_{ij}^{(k)}$ is the number of paths connecting V_i and V_j with length of k. We define an associate degree to measure their connectivity.

$$S_{ij}=k_1a_{ij}^{(1)}+k_2a_{ij}^{(2)}+...+k_pa_{ij}^{(p)} \quad (2)$$

Here, p is a positive constant integer, k_i (i=1,2,..,p) are coefficients. We set the pheromone with the same value and set $\eta_{ij}=S_{ij}$. In our experiments, we set k_p=1, k_i=2k_{i+1}, and $p=n.log_{10}n$.

3.3 Pheromone updating

Line 5 of algorithm ACODCS updates the pheromone value on each arc according to the formulas as follows:

$$\tau_{ij}(t+1)=\rho\cdot\tau_{ij}(t)+\Delta\tau_{ij}(t) \quad (3)$$

Here,

$$\Delta\tau_{ij}(t)=\sum_{k=1}^{m}\Delta\tau_{ij}^{k}(t) \quad (4)$$

Here, $\Delta\tau_{ij}^{k}(t)$ can be calculated as follows: if in the k-th solution S_k, v_i and v_j are not in the same community, then $\Delta\tau_{ij}^{k}(t)=0$. Otherwise $\Delta\tau_{ij}^{k}(t)=C.Q(S_k)$, here C is a constant and $Q(S_k)$ is the quality of solution S_k.

Obviously, the more ants assign V_i and V_j in the same community, the more increment of pheromone in the arc E_{ij} has, and the higher probability the ants select the arc E_{ij} in the next iteration. This forms a positive feedback of the pheromone system.

3.4 Modularity

To evaluate the quality of the solutions obtained by the ants, we use the modularity measure proposed by Newman [15]. Modularity, which is also called the strength of a network, counts the number of links between all pairs of nodes belong to the same community and compares it to the expected number of such links for an equivalent random graph in which the degree of all nodes has been left unchanged. Let $S=\{C_1,\ldots,C_k\}$ be a solution where $C_1,\ldots,C_k$ are k clusters in S, we define its modularity $Q(S)$ as:

$$Q(S)=\frac{1}{2m}\sum_{l=1}^{k}\sum_{(i,j)\in C_l}\left[a_{ij}-\frac{\deg(v_i)\times\deg(v_j)}{2m}\right] \quad (5)$$

Here, $2m=\sum_{i=1}^{n}\deg(Vi)=\sum_{i,j}a_{ij}$ which is the sum of the degree of all vertexes. With modularity, the detection of community structure becomes an optimization problem of the modularity among all the possible solutions. The goal of the ant's exploration is to find a solution with the highest modularity.

3.5 Termination conditions

Step 7 in the algorithm ceases the iterations according to a certain termination condition. We stop the iterations when the clusters obtained in adjacent iterations tend to stabilize. In addition, we also set up a threshold Nc, which is the maximum number of iterations. The iterations should be ended as well when the number of iterations goes beyond Nc.

4 Experimental results

In this section, we empirically demonstrate the effectiveness of our proposed algorithm ACODCS on synthetic benchmarks networks and real world networks. All experiments have been run on Pentium IV, Windows XP, P1.7G, and visualize the results on Matlab 6.0.

4.1 Tests on synthetic benchmarks networks

We adopt the benchmark proposed by Lancichinetti et al. [37].This benchmark provides networks with different nodes degrees and community sizes, which are common characteristics in real world networks. Many parameters are used to specify the generated networks, such as the nuber of nodes n, the average node degree Z, the average community size n_a, *number of communities* k. In our experiments, we set n=128, Z=16, n_a=32, k=4. We have generated a large number of random graphs with known community structure. In the graphs, each vertex has on average Z_{in} edges connecting it to the nodes in the same community and Z_{out} edges to the nodes in other communities. The values of Z_{in} and Z_{out} should be chosen to coincide with the expected degree of each vertex Z. We set $Z_{in}+Z_{out}$ =16. We test our algorithm and compare its performance with that of the Fast algorithm[21], the GN algorithm[24] and the Hierarchical clustering algorithm [32]. Figure 4 shows the percentage of the nodes assigned to the correct communities in the solution by our algorithm and other three algorithms under different values of parameter Z_{out} .

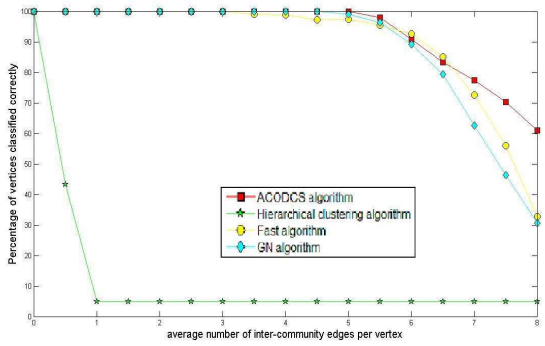


Figure 4 The percentage of vertices correctly identified by four algorithms

From Figure.4 we can see that our algorithm ACODCS has the highest percentage of correctly assigned nodes under all different values of parameter Z_{out}. For instance, when Z_{out}=6, ACODCS can correctly identify more than 90% of vertices. Especially for the values of Z_{out} from 5 to 8, the percentage is much higher than other three algorithms. For instance, when Z_{out}=8, ACODCS can correctly identify more than 60% of vertices, while other three algorithms can only identify 31%,30% and 4% respectively.

4.2 Zachary's karate club network

The network of "Zachary's Karate Club" proposed by Zachary[33] illustrates the pattern of friendships between 34 members of a karate club at a US university in 1970s. Because shortly after the observation and construction of the network, the club split into two clubs as a result of an internal dispute, this example is of particular interest. Recently, this network has become something of a standard benchmark for testing community detection algorithms. Vertices in the graph represent the members in the club, while the edges represent relationship between the two people they connect.

If the network is partitioned just as the way the Karate club broke up into two separate clubs over a conflict, the modularity of this solution is Q=0.3715. This network is also discussed in [17] that four communities were found by the authors, and the Q is 0.4197 that is the maximum value ever found of modularity for this network. Figure 5 shows the community structure detected by our algorithm for Zachary's Karate Club network. In the figure, each community is shaded with a different color.

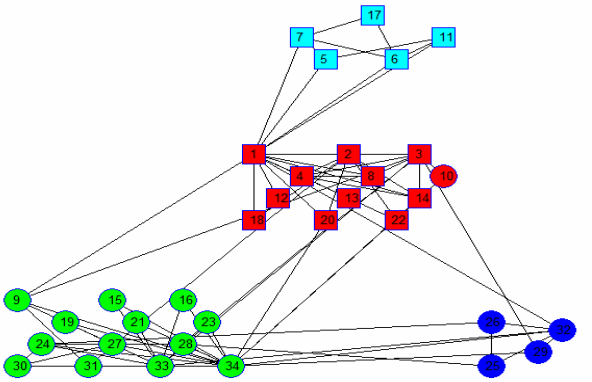


Figure 5 The community structure detected by our algorithm for Zachary's Karate Club network. Each community is shaded with a different color.

As shown in Figure 5, four communities are detected by our method, with modularity value Q =0.4165 which is much higher than the original solution and is very close to the maximum.

4.3 College football

The network of college football represents an American college football games between Division I colleges during the regular season in fall 2000. Vertices in the graph represent the teams, and edges represent regular season games between the two teams they connect. The teams are organized into several communities, and each of those community consists of 8 to 12 teams. Usually, more games are played within a community than across the communities. Modularity of the solution reflecting the real community partition is Q=0.5540.

Figure 6 shows the community structure detected by our algorithm for the college football network. In the figure, each community is shaded with a different color. As shown in Figure 6, eleven communities are detected by our method, and the modularity of the solution is with Q=0.6031 which is much higher than the original solution. This network is also discussed in [17] that ten communities were found by the authors, and the modularity value is Q=0.6046 which is the maximum value found so far. We can see that the modularity value of the solution obtained by our algorithm is very close to the maximum.

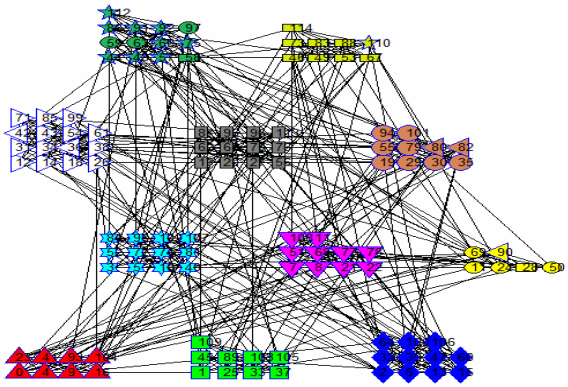


Figure 6 The community structure detected by our algorithm for College football network. Each community is shaded with a different color.

4.4 Books on American politics

The network of Books on American politics representing the frequently co-purchased books was compiled by Krebs. Vertices in the graph represent books on American politics bought from amazon.com, and edges represent frequently co-purchases of the two books they connect. Originally, the books are divided into three communities, the modularity is Q=0.4149. This network is also discussed in [17] that five communities were detected, and the modularity of this solution is Q=0.5272 which is the maximum value modularity for this network found so far.

Figure 7 shows the community structure detected by our algorithm for the Books on American politics network. As shown in Figure 7, five communities are detected by our method, and the modularity of the solution is with Q=0.5262, which is much higher than the original solution and is very close to the maximum.

For the three networks mentioned above, Figure 8 shows the comparison of the modularity values of the original solution, the solution by our algorithm and solution with the maximum modularity found so far. From the figure, we see that the modularities of the solutions obtained by our algorithm are much higher than the original solution and very close to the maximum.

4.5 Dolphin's associations

The network of Dolphin's associations was presented by D.Lusseau and K.Schneider[34]. Vertexes in the network represent 62 dolphins living in Doubtful Sound, New Zealand, while the edges reflects the ties between dolphin pairs. It is believed that dolphins with statistically significant frequent ties form an association.

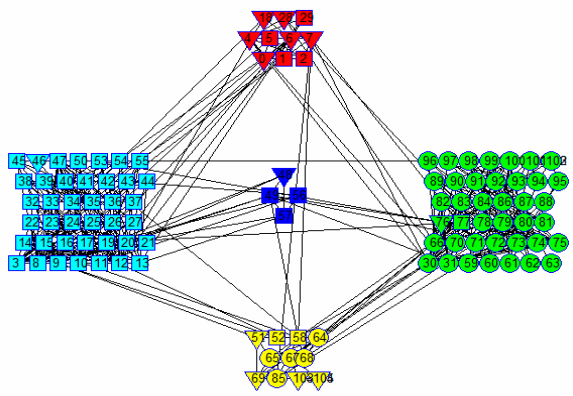


Figure 7. The community structure detected by our algorithm for Books on American politics. Each community is shaded with a different color.

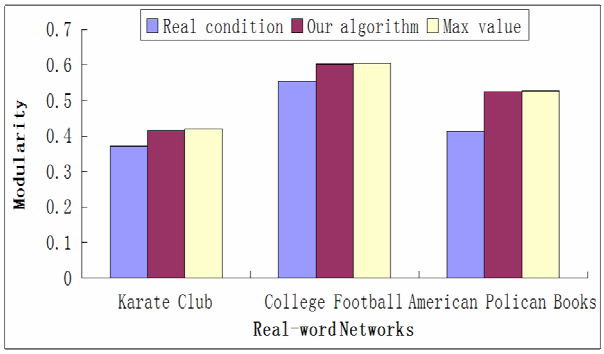


Figure 8 Comparison of the modularities of three solutions.

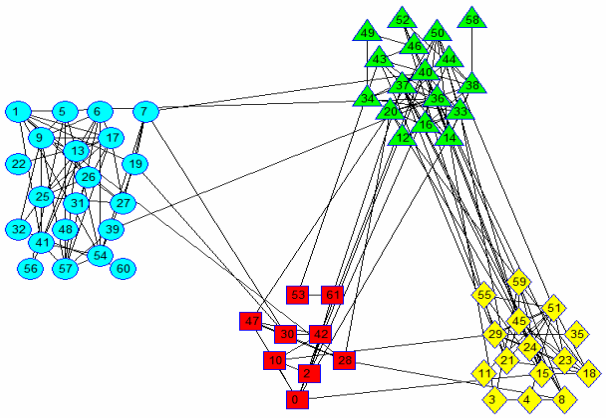


Figure 9 The community structure detected by our algorithm for Dolphin's associations. Each community is shaded with a different color.

Figure 9 shows the community structure detected by our algorithm for the Dolphin's associations network. As shown in Figure 9, four communities are detected by our method, with modularity value Q=0.5628. This network is also detected by D.B. Chen, Y. Fu and M.S. Shang[36], modularity of their solution is Q=0.3286. Moreover， T.S. Evans and R. Lambiotte [35] obtained a solution on this network with modularity value Q=0.5481. We can see that our algorithm obtains the solution with the highest modularity.

From the experiments, we can see that our algorithm can achieve high quality results on both synthetic and real-world networks.

5 Conclusions

In this paper, we present a new algorithm based on ant colony optimization to detect community structure in networks. In the algorithm, artificial ants are used to travel on a logical digraph to construct solutions of community detection. Each ants chooses its path according to the pheromone and heuristic information on each path. We define an associate degree and use it as the heuristic information. The quality of solution obtained by each ant is measured by its modularity. The algorithm is tested on synthetic and real world networks. The experimental results show that our algorithm can achieve higher quality results than other methods.

6 Acknowledgments

This research was supported in part by the Chinese National Natural Science Foundation under grant No. 61070047 and 61070133, Natural Science the Scientific Research Foundation for Graduated Students in Jiangsu Province.

7 References

[1] M.E.J Newman, The structure and function of complex networks[J], SIAM Rev. 45 (2003) 16-256.

[2] S.H. Strogatz, Exploring complex networks[J], Nature 410 (2001) 268-276.

[3] Wasserman&Faust. Social Network Analysis (Cambridge Univ. Press, Cambridge, U.K) (1994).

[4] Scott, J. Social Network Analysis: A Handbook [M](Sage, London), 2nd Ed,(2000).

[5] Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks [J].Nature (London) 393, 440–442.(1998).

[6] Amaral, L. A. N., Scala, A., Barthelemy, M. & Stanley H. E., Classes of small-world networks [C]. Proc. Natl. Acad. Sci. USA 97, 11149–11152. (2000).

[7] Williams, R. J. & Martinez N. D. , Food-web structure and network theory: The role of connectance and size[J]. Nature (London) 404, 180–183(2000) .

[8] Faloutsos M,Faloutsos P,Faloutsos C. On power-law relationships of the internet topology. Communication Review[C], In: Proceedings of ACM SIGCOMM, ,29251-262(1999).

[9] Albert, R., J., H., and Barabasi, A.-L. Diameter of the World-Wide Web[J], Nature 401, pp. 130–131(1999).

[10] Newman, M. E. J. The structure of scientific collaboration networks[C], (2001) Proc. Natl. Acad. Sci. USA 98, 404–409.

[11] Newman M E J. Fast Algorithm for Detecting Community Structure in Networks[J] . Phys Rev E, 2004, 69(6) : 066133.

[12] Clauset A. Finding Local Community Structure in Networks [J] . Phys Rev E, 2005, 72(2) : 026132.

[13] Schuetz P, Caflisch A. Multi step Greedy Algorithm Identifies Community Structure in Real world and Computer generated Networks[J] . Phys Rev. E, 2008, 78(2) : 026112.

[14] Girvan M, Newman M E J. Community Structure in Social and Biological Networks[J] . P Natl Acad Sci USA, 2002, 99 (12) : 7821- 7826.

[15] Newman M E J. Modularity and Community Structure in Networks [J] . Proc Natl Acad Sci USA, 2006, 103(23) : 8577- 8582.

[16] Duch J, Arenas A. Community Detection in Complex Networks Using Extremal Optimization [J] . Phys Rev E, 2005, 72 (2) : 027104.

[17] Agarwal G, Kempe D. Modularity maximizing Graph Communities Via Mathematical Programming[J]. The European Physical Journal B, Condensed Matter and Complex Systems, 2008, 66(3) : 409- 418.

[18] Ruan JH , Zhang W X. Identifying Network Communities with a High Resolution[J] . Phys Rev E, 2008, 77(1) : 016104.

[19] Li Z, Zhang S, Wang R S, et al. Quantitative Function for Community Detection [J] . Phys. Rev. E. Stat. Nonlin Soft Matter Phys, 2008, 77(3) : 036109.

[20] Donetti L, Munoz M. Detecting Network Communities: A New Systematic and Efficient Algorithm [J] . Journal of Statistical Mechanics, 2004, P10012.

[21] Capocci A, Servedio V D P, Caldarelli G, et al. Detecting Communities in Large Networks [J] . Physica A: Statistical Mechanics and Its Applications, 2005, 352(2-4) : 669- 676.

[22] Rosvall M, Bergstrom C T. An Information theoretic Framework for Resolving Community Structure in Complex Networks [J] . P Nat l Acad Sci USA, 2007, 104(18) : 7327- 7331.

[23] Rosvall M, Bergstrom C T. Maps of Random Walks on Complex Networks Reveal Community Structure [J] . Proc Natl Acad Sci USA, 2008, 105(4) : 1118- 1123.

[24] Raghavan U N, Albert R, Kumara S. Near Linear Time Algorithm to Detect Community Structures in Large scale Networks[J] . Phys Rev E, 2007, 76(3) : 036106.

[25] Leung I, Hui P, Lio P, et al. Towards Real time Community Detection in Large Networks [J] . Physical Review E, 2009, 79 (6) : 66107.

[26] Palla G, Derenyi I, Farkas I, et al. Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society [J] . Nature, 2005, 435(7043) : 814- 818.

[27] Kumpula J M, Kivel M, Kaski K, et al. Sequential Algorithm for Fast Clique Percolation [J] . Physical Review E, 2008, 78, (2) : 026109.

[28] Zhang S, Wang R, Zhang X. Identification of Overlapping Community Structure in Complex Networks Using Fuzzy C-means Clustering [J] . Physica A: Statistical Mechanics and its Applications, 2007, 374(1) : 483- 490.

[29] Ding F, Luo Z, Shi J, et al. Overlapping Community Detection By Kernel based Fuzzy Affinity Propagation [C], Proceedings of the 2nd International Workshop on Intelligent Systems and Applications(ISA& 2010) , Wuhan, China, 2010.

[30] Zhang S, Wang R S, Zhang X S. Uncovering Fuzzy Community Structure in Complex Networks[J] . Phys Rev E, 2007, 76(4) : 046103.

[31] Zarei M, Izadi D, Samani K A. Detecting Overlapping Community Structure of Networks Based on Vertex-vertex Correlations [J] . Journal of Statistical Mechanics: Theory and Experiment, 2009, P11013.

[32] Lancichinetti A, Fortunato S, Kert esz J. Detecting the Overlapping and Hierarchical Community Structure in Complex Networks [J] . New Journal of Physics, 2009, 11: 033015.

[33] W.W. Zachary, An information flow model for conflict and fission in small groups[J], J. Anth. Res. 33 (1977) 452-473.

[34] D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten, S.M. Dawson, The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations[J], Behav. Ecol. Sociobiol. 54 396-405 (2003) .

[35] T.S. Evans, R. Lambiotte, Line graphs, link partitions, and overlapping communities[J], Phys. Rev. E 80 (2009) 016105.

[36] D.B. Chen, Y. Fu, M.S. Shang, An efficient algorithm for overlapping community detection in complex networks[C], in: 2009 Proc. Global Congress on Intelligent Systems, 19-21 May, Xiamen China, pp. 244-247.

[37] Lancichinetti A, Fortunato S, Radicchi, Benchmark graphs for testing community detecting algorithms[J], Phy. Rev. E,78(4):046110,2008

Ranking of telephone conversations on the basis of the client's emotional attitude

A. Kocsor[1], K. Kovács[1], A. Bódogh[2] and I. Fehér[2]

[1]Applied Intelligence Research Nonprofit Ltd., Szeged, Hungary
[2]Research & Development Unit, Xdroid Ltd., Szeged, Hungary

Abstract – *An innovative, mostly language-independent technology has been created for customer call centers that ranks telephone conversations on the basis of the probability of their business outcomes. The system stands out with its new methods of optimizing samples to a ranked list and with its mechanical learning process designed specifically for this purpose; which also proves to be original as an approach. The team has successfully tested the system at a company selling timeshare holidays with remarkable results. The list ranked by the system helped them reach their potential clients more quickly and efficiently. In the experimental set of conversations they observed an improvement of 24% - 72% depending on the number of actual recalls. Calls made in accordance with the top of the ranked list resulted in considerable profits and faster sales process.*

Keywords: machine learning, artificial intelligence, speech analytics, emotion detection, data mining.

1 Introduction

Speech studies have recently made significant strides due to the development of the speed, memory and storage of computers. Different branches of speech-related research can be inventoried by considering which part of speech they aim to recognize.

1.1 Background

1.1.1 Speaker recognition

In the event of multiple participants in a dialog, it may be necessary to distinguish and segment each speaker. This task usually appears as part of the pre-processing section of another speech-oriented task. During human speech speakers often communicate simultaneously further complicating the situation. This has to be recognized as well and these segments must be annotated separately [1].

1.1.2 Speaker search

The goal is to find a specific speaker in a "crowded" sound sample on the basis of a voice pattern. The output is the labeling of segments in which the speaker in question is most likely to be heard.

1.1.3 Speaker identification

Here, the task is the automatic, real-time identification of the speaker with the help of a stored sound sample. It is necessary typically due to eligibility and security considerations such as in the case of access control systems or during phone or online administration [2][3].

1.1.4 Speech recognition

The task is to identify spoken language sequences. It can be keyword recognition, in which case the goal is to find the words of a limited list of words in a dense sound sample (keyword search). Or, it can be non-limited word recognition, which is actually "real" speech recognition, which represents domain-independent complete recognition [4].

1.1.5 Emotion recognition

The goal here is the detection of predefined human emotions. Since emotions might surface in language elements as well, keyword recognition can increase the efficiency of emotion recognition [5].

1.1.6 Language recognition

The task is to identify the language of the speakers. It is necessary for the search domain (the sound sample) to contain conversation in multiple languages. This task is also essential in the case of real-time speech recognition [6].

1.1.7 Noise recognition

The separation of non-speech-related sequences, and then the classification of sequences that are considered "noise" (background music, traffic sounds, noise generated by the speakers themselves, etc.). They usually are background noises and are perceived simultaneously with human speech, which makes the task rather complicated [7][8].

1.1.8 Special sound recognition

The identification of special sound types (hoarse, nasal or sniffling voice, whispering, humming, singing, etc.) challenges researchers considerably, as well [9].

2 Current research

Speech technology applications are characteristically not unipolar because they integrate more of the aforementioned branches. Under this research the participant companies have developed such a multipolar application. The team set out to design an original, language-independent technology that ranks clients on the basis of customer service telephone recordings.
How can companies profit from such ranked lists of clients?
A typical use case is when a company that employs agents would like to market their products more effectively. For example, when agents are selling timeshare holiday, and are confronted with a list of hundreds of clients who have already been contacted about the product in question. Nevertheless, the potential buyers have not yet given a firm yes or no. Their initial conversations have been recorded. But how to proceed?

- Should the agents listen to some or all the conversations again and intuitively (or randomly) select telephone numbers to redial?
- Shall they call everyone on the list once again?

The last is quite expensive and takes significant time. The first, on the other hand, is arbitrary and, considering its efficiency, unpredictable. How shall agents decide which numbers should be redialed and which client will end up accepting the offer?

This is what our research technology aims to solve with the combination of the following primary pillars:

- Speech processing that considers many aspects of speech analytics and relies heavily on emotion recognition,
- Ranking process optimalization, using a new approach for machine learning.

In the following chapters we give details about the achievements.

2.1 The task of recognizing emotions

Researchers of the recognition of human emotions can examine a number of sources in order to achieve their goals:

- visual information: gesticulation and facial expressions,
- physiological parameters: brain activity, change of states in the limbic system, perspiration, etc.,
- auditory information: the characteristics of the speech process.

The purpose of the present research is to detect emotions on the basis of recorded conversations. Consequently, there are no visual or physiological cues, only audible files are available. Furthermore, a restricted acoustic sign is transmitted via phone, therefore a significant part of the live speech bandwidth is not available. Nevertheless, one can determine the emotional state of the speaker more or less precisely even by this limited channel. If humans are capable of discerning emotions via phone, then what limits a computer from doing the same? According to a recent study, even three-month old infants are able to detect negative intonations in human speech [10].

In spite of this, the task is not simple. Its complexity is revealed by a number of other factors. On the one hand, the average accuracy of auditory emotion recognition (i.e. relying only on aural content) of humans is only about sixty percent [11]. On the other hand, despite extensive research, professionals cannot seem to agree which acoustic features may help the most in precisely identifing or distinguishing human emotions [12].

2.2 Parameters influencing efficiency of emotion recognition

The accuracy and detection rate of emotion recognition is affected by numerous factors. The major influencers are:

- Quality of sound sample:
  1. environmental noise and its subtypes:
     a. constant, low-level background noise;
     b. sudden, short, typically loud noise (ambient, or speaker-originated e.g. breathing, coughing, throat clearing);
  2. quality of digitalization: bandwidth, sample size, quality of interconnected devices (limited frequency: loss of information);
  3. technical attributes of transmission channel (limited sound track: loss of information).

- Quality and technical details of local speech databases:
  1. not telephone-related: MRBA, BABEL, SPECO, BEA: 16-20 kHz, 16 bit;
  2. telephone-related: MTBA, SpeechDat-E: 8kHz, 16 bit.

- Artificial vs. spontaneous speech:
  1. *artificial:* recited or read (text or list of words); purpose: clear recording, instances of frequent sound clusters (learning to recognize, building blocks to synthesis); emotions are not spontaneous;
  2. *semi-spontaneous:* controlled conversation: predefined topic or interview;
  3. *spontaneous:* observers' paradox: to observe participants while ensuring true spontaneity.

Emotion recognition is simpler using a database containing spontaneous speech, however, the accuracy level of these methods is significantly less than that of artificial speech.

- Size of learning database:
  1. common elements in sample deviations;
  2. precision of recognition does not increase beyond a certain sampling size and a given number of speakers.

- Speaker-dependent or speaker-independent recognition: are learning and recognition adapted to the speaker?
  1. adapted to the speaker: this method provides much better results, but it cannot be applied online because it takes too much computational time and it requires speaker-dependent learning [13][14];
  2. speaker-independent: universal because it does not depend on personal features (normalization), but its accuracy of recognition is lower [15].

2.2.1 Number of emotions to be distinguished and recognized

Along with verbal content, supra-segmental and non-verbal information also belongs to communication. These extra components might supplement, modulate and sometimes even utterly modify the content of the speech. The situation is further complicated by the fact that emotional involvement can be expressed explicitly in the verbal content, although non-verbal signs can usually be detected in such cases as well.

Another problematic aspect of emotion recognition is that the actual form of an emotion varies considerably in quality and intensity between speakers of different genders and ethnic backgrounds. Moreover, the very same person can express a feeling differently at different times. The greater number of emotions that need to be detected, the less accurate our recognition will be. In other words, if we distinguish a large number of distinct types of emotion, an inaccurate detection is more likely to occur. This means that if we dedicate fewer classes for emotions, the more reliable the designation can be.

Possible classifications with an ample number of classes:

N: normal, neutral (Neutral);
H: harmonious, happy, satisfied (Happy);
D: angry, displeased, vengeful, dissatisfied (Displeased);
S: sad, depressed, disappointed (Sad);
F: afraid, nervous, trembling voice (aFraid);
X: undistinguishable, or other (eXtra).

A possible reduced classification, in which negative emotions (D, S, F) are merged:

N: normal, neutral;
H: happy, satisfied;
D: displeased.

2.3 Recognition time

It depends on the task itself whether a real-time recognition is necessary or one that is offline where later analysis is also possible. Online recognition is less accurate for the following reasons: First, recognition time is quite short, and second, in the case of real-time recognition, the left side (the past) of the analyzed segments are available for analysis.

2.4 Efficiency of the recognition system

The efficiency and robustness of the system can be increased if expert rules are added to the statistics-based classification. Expert rules linked to emotions, which were created on the basis of the pre-evaluated database:

Afraid:
- trembling, quivering;
- silent;
- without characteristic intonation (babbling);
- floating, wavering;
- interrupted;
- weeping.

Displeased:
- stress is not released on the final syllables of a sentence;
- rising intonation, stress on final syllables;
- final syllable is lengthened;
- first syllable is very high-pitched and emphatic;
- rhythm: either articulated, slow and threatening (indicating that it should be more vehement, but it is being held back), or quick (in the midst of releasing anger);
- loud, exasperated.

Happy:
- “jovial speech”;
- boastful: intonation varied rhytm;
- clauses have a lengthened, floating ending;
- enthusiastic.

Sad:
- quiet;
- no intonation, except for a rise in pitch towards the end;
- slow, measured;
- specific sounds are stressed at the beginning of negative words (a language-specific pattern).

2.5 Ranking process

The ranking process is a filtering method that significantly helps to acquire new clients and keep old ones, that is, it potentially increases business profitability. From a scientific point of view, ranking takes place through computer learning on the basis of the thorough analysis of telephone conversations. The purpose of analysis is to distinguish the different speech dynamics features (speaker distinction, communication features, keyword recognition, emotion detection, psycholinguistic characteristics) and assign an value to them.
Now we define the Success Index (SI). SI of number 1 is assigned to each closed customer conversation, in the event that the sale was successful, and if it is 0, it was not. If there is an adequate number of (lifeline,SI) pairs, then the correlation between the success index and the speech-related features (the emotions and psycholinguistic characteristics of the lifeline along with how they changed through time) can be examined. The tendencies and correlations found therefore can provide

predictions regarding the success indexes of the lifelines in progress. In other words, it can be anticipated which lifelines are worth continuing, which are the ones that will result in a new customer.

The learning input of the ranking process is therefore (lifeline,SI) pairs, and its output is a ranking model that ranks clients' lifelines by their predicted SI. The numeric value of the indices is refined by determining them in the [0, 1] interval. Ranking therefore puts those to the head of the list that show as worthwhile to call. So the company should consider to call only the top of the list henceforth. When the model (the ranking engine) operates effectively, the company will lose very few potential customers (if any) if it decides to ignore the coldest clients at bottom of the list.

A further innovative aspect of the system is that a computer learning method that has been designated for this particular task and optimizes the ranking engine (the ranking process). The operational efficiency index of the ranking engine is the lift measure. Lift is a measure of the performance of a model at the predictive ranking as having an enhanced response (with respect to the population as a whole), measured against a random choice model. Lift is the ratio of these values: model response divided by average response. Let ModelResponse be the percent of succesful target response in the first percent (p) of the list ranked by decreasing score of a given model. Let AverageResponse be the percent of succesful responses in the first p of the list using random choice selection. The lift can be defined as

$$\text{Lift}(p) = \frac{\text{ModelResponse}\ (p)}{\text{AverageResponse}(p)} \qquad (1)$$

Computer learning methods are traditionally optimized for accuracy. In contrast, this new approach optimizes the learning system for the aforementioned lift measure. This is crucial because ranking customers to determine increased sales does not necessarily correlate with greater accuracy. A relatively high accuracy might be theoretically possible with a decreased lift measure, however, it results in a model that is irrelevant at bringing more sales and increased profit.

2.6 General characteristics of the system

The developed system features multi-lingual core algorithms, meant for the international market and it prioritizes customers on the basis of telephone conversations. The phone conversation of the each customer are chronologically ordered (as the "lifeline" of the client) and the system evaluates all conversations in a client's lifeline. The SI is assigned to each lifeline that suggests whether the client was worth acquiring or retaining. The emotions detectable from the client's speech, certain psycholinguistic characteristics and their changes through time are all actively linked with the SI. Researchers found that on the basis of multiple features of speech dynamics it can be quite reliably predicted which open lifelines are worthwhile to continue, resulting in new customers and business. In other words, it can be anticipated which clients are worth calling back and which clients are not worth further contact.

One of the most significant scientific innovations of the system is that it applies computer learning algorithms in a unique calculative environment. Usually, several high-capacity processors are required because efficient speech processing requires resource-intensive complex calculations. During this research these calculations have been realized with the application of massive parallel processing based on GPU (Graphics Processing Unit). For several years the usage of GPU has been expanding within international research projects. Our research team has been successfully applying it within this research. GPU, with its significant calculative capacity, has enabled us to handle multiple layer of speech processing efficiently. This techology is quite adequate at examining numerous correlations between elements of communication. In another words, we can distinguish more features that are simultaneously present on multiple layers using its computational strength. It is technically relevant that GPU can be used in an ordinary PC configuration, and thus no super computer is necessary to reach the desired result.

Another essential innovation is in connection with the distinction of speech dynamics features. This distinction is the identification of relevant elements in the communication between customer and agent (e.g. change of speakers, length of speech stretches, keywords mentioned, speakers' emotions, psycholinguistic characteristics). The processing of these elements (features) requires an approach of more perspectives, presupposing partly consecutive, partly simultaneous processes. Some of the features are distinguised automatically and others semi-automatically, and others applying expert rules. When one assigns numbers to the features, the system is capable to recognize and learn correlations and tendencies within conversations using computer learning techniques. Independently of the length of a conversation, the system always distinguishes a constant number of speech dynamic elements. In other words, the same features need to be definable in the shortest speech stretches and must also remain in longer dialogs. Distinct features correlate differently with the SI. In order to properly define the SI, we must extract a constant number of features from each conversation.

2.7 Results

Validation tests were conducted at the customer service center of a company dealing with timeshare holidays. Customer service had to call potential customers to invite them for an onsite sales event. Customers who did not refuse the offer out right were called back after a while. Calls were considered successful, if the customer accepted the invitation, and unsuccessful, if he or she refused. With the help of a client-tailored ranking model, based on success probability we ranked completed calls following the first conversation. Utilizing the upper percentile of our ranked list, the company anticipated a higher success rate upon completion the second call. Learning of the model was conducted on a database of more than 8,000 labeled conversations with 408 positive cases. During testing, the samples were separated, 60% used

for learning, 40% used for testing. Then we measured the ranked list containing 3,200 conversations, out of which 167 were known to be positive. The results of the ranking can be seen in Table 1:

Table 1 – Ranking results

Percentage of calls	Unranked list (No. of attendees)	Ranked list : (No. of attendees)	Lift measure (efficiency multiplicator)
1%	1.67	11.7	7.01
2%	3.33	15.57	4.67
5%	8.33	54.5	6.54
10%	16.67	70.83	4.25
25%	41.67	82.92	1.99
50%	83.33	142.5	1.71
75%	125	155	1.24
90%	150	159	1.06
100%	166.67	166.67	1

According to Table 1, at the top 10% level, sales results increased 4 times more than compared to random selection. Moreover, at the top 1% level, the company experienced a sales boost of more than 7 times.
In other words, calling 50% of the numbers formerly made it possible to reach 50% of the participants (which was 83 clients), with the use of ranking, it became 85% (143 clients), which represents a 72% improvement. Formerly, with 75% of the calls, 75% (125 clients) of the clients were successfully invited, and with ranking, it became 93% (155 clients) of all the participants, which represents a 24% improvement. All in all, one can assume that calls made in accordance with the ranked list can result in considerable profits.

2.8 Conclusions

The system is partly language-independent which enables it to be quickly adapted all the world's major languages. Furthermore, with the help of the system, the success of sales agents can be meaningfully evaluated, which means that a (lifeline,SI) pair can be assigned to them as well. In the long run, it enables clients to hire more efficient agents. The method can be effectively adapted for use with any other customer business processes involving phone communication and increase its efficiency. Companies which market their products and services via phone can utilize the system in order to acquire as well as retain clients more efficiently, to optimize customer retention and to monitor their own employees' job efficiency.

3 Acknowledgement

Research supported by the European Union in cooperation with the Hungarian Economic Development Operational Program under grant GOP-1.1.1-09/1-2009-0015.

4 References

[1] Docío-Fernández, L. & García-Mateo, C.: "Speaker Segmentation". Encyclopedia of Biometrics. pp. 1277-1284, 2009.

[2] Keshet, J. & Bengio, S.: "Automatic speech and speaker recognition". John Wiley and Sons, 2009.

[3] Jin, Q.: "Robust Speaker Recognition". Language Technologies Institute School of Computer Science, Carnegie Mellon University, 2007.

[4] Becchetti, C. & Ricotti, L. P.: "Speech Recognition: Theory and C++ Implementation". Viley, 1999.

[5] Hozian, V. & Kacic, Z.: "Context-Independent Multilingual Emotion Recognition from Speech Signals"; International Journal of Speech Technology 6. pp. 311-320, 2003.

[6] Hermansky, H. & Adami, A. G.: "Segmentation of Speech for Speaker and Language Recognition" ; Eighth European Conference on Speech Communication and Technology. pp. 841-844, 2003.

[7] Kubica, J. & Moore, A.: "Probabilistic Noise Identification and Data Cleaning". Robotics Institute Carnegie Mellon University, 2002.

[8] Couvreur, L. & Laniray, M.: "Automatic Noise Recognition in Urban Environments Based on Artificial Neural Networks and Hidden Markov Models"; Proceedings of InterNoise, Prague, Czech Republic, 2004.

[9] Salhi, L., Mourad, T. & Cherif, A. : "Voice Disorders Identification Using Multilayer Neural Network" ; The International Arab Journal of Information Technology, Vol. 7, No. 2, pp. 177-185, 2010.

[10] Blasi, A., Mercure, E., Lloyd-Fox, S., Thomson, A., Brammer, M., Sauter, D., Deeley, Q., Barker, G. J., Renvall, V., Deoni, S., Gasston, D., Williams, S. C. R., Johnson, M. H., Simmons, A., Murphy, D. G. M.: "Early Specialization for Voice and Emotion Processing in the Infant Brain"; Current Biology, Volume 21, Issue 14, pp. 1220-1224, 2011.

[11] Tóth Sz. L., Sztahó D., Vicsi K.: "Speech Emotion Perception by Human and Machine, Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction". Springer-Verlag Berlin, Heidelberg, pp. 213-224, 2008.

[12] Laukka, P.: "Vocal expression of emotion: Discrete-emotions and dimensional accounts" ; Comprehensive summaries of Uppsala dissertations from the faculty of social sciences, No. 141. Uppsala, Sweden: Acta Universitatis Upsaliensis. pp. 1-80, 2004.

[13] Kolár, J., Liu, Y. & Shriberg, E.: "Speaker Adaptation of Language Models for Automatic Dialog Act Segmentation of Meetings" ; In Proc. Interspeech, pp. 1621-1624, Antwerp. 2007.

[14] J. Yamagishi, J., Nose, T., Zen, H., Ling Z. H., Toda, T., Tokuda, K., King, S., Renals, S.: "Robust speaker-adaptive HMM-based text-to-speech synthesis". Audio, Speech, and Language Processing, IEEE Transactions vol. 17 (nr. 6), pp. 1208-1230, 2009.

[15] Heeyoul, C. H., Gutierrez-Osuna, R., Choi, S. & Choe, Y. : "Kernel Oriented Discriminant Analysis for Speaker-Independent Phoneme Spaces", Pattern Recognition, ICPR 19th International Conference, 2008.

A Topic-Map-Based Framework for Decision Information Systems

STMicroelectronics' case study

S. Bouzid[1,2], **C. Cauvet**[1], **and J. Pinaton**[2]
[1]Laboratory for Systems and Information Sciences (LSIS), Marseille, France
[2]STMicroelectronics, Rousset, France

Abstract *- With the increase of company components for data processing, calculation and reporting -known as Decision Information Systems (DIS)-, many companies seek today to enhance the sharing and the retrieval of such components among end-users to have effective returns on investment. The specificity of these components is that they highly support business requirements in a company with a set of indicators. We propose in this paper a methodological framework based on the Topic Map standard to integrate business semantic in a DIS. The aim of this work is to improve the sharing and the retrieval of DIS components among end-users in a company. The characteristic of this framework is that it provides a requirement-oriented description to DIS components and a methodology to integrate this semantic in an iterative approach.*

Keywords: Decision Information System; Semantic Description; Topic Maps; Business Requirements

1 Introduction

A Decision Information System (DIS) is a component of Information System aiming at processing large amount of data in order to identify business tendencies and to support business-processes' performance. The components of such a system consist of a set of data sources like company data bases and data warehouses supported with a set of decision-support applications for data collection, transformation and reporting. These components are intended for different profiles of end-users (i.e. managers, executives and operatives) enabling them to produce and retrieve aggregate of measures called indicators (figure 1).

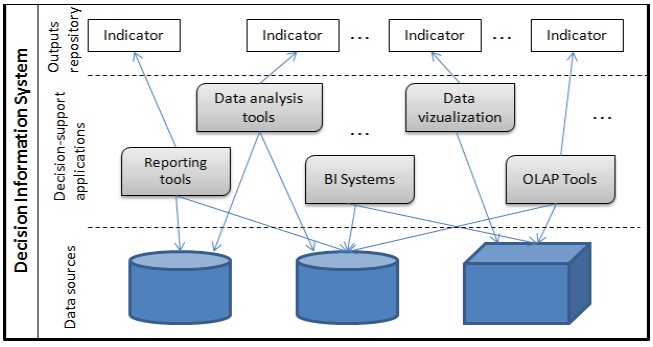


Figure 1. Components of a DIS

DIS have acquired a wide importance in the business world nowadays because they are involved in many business purposes. Examples include the control of business processes such as manufacturing processes, healthcare processes, the management control, etc. The key success of establishing DIS in companies consists not only in how to design such systems according to business expectations, but also in how to efficiently use them by the end-users. In fact, many companies adopt DIS to support their core business for two purposes: (a) to enable end-users to use same data and consistent data sources to produce accurate indicators, and (b) to enable the sharing and the retrieval of these indicators by end-users so to avoid redundancies and the waste of time in the development of same needs. However, the second purpose is not really enough tackled by researchers comparing to the first one. Most research works devoted to DIS focus on approaches to design data-sources models and decision-support components [1] [2] [3] [4]. Even if some of these approaches are successful [5], users still complain about mismatching with their needs mainly because there are heterogeneous DIS components in the company and users don't know how to retrieve existing indicators. Indeed, an indicator can be created with different types of components such as by querying with OLAP tools and BI systems or by using complex components for statistical data analysis. It is important that the users who have same needs can share and reuse same components. This process needs to be efficiently handled.

To that aim, we propose to use semantic description techniques, widely used by the scientific community to address sharing and retrieval issues. The challenge in using such techniques is to identify what kind of semantic knowledge must be represented and how to implement it in an existing system. We propose in this paper a methodological framework to integrate business semantic in a DIS using the Topic Map standard which is an ISO semantic web standard. This semantic allows retrieving DIS components by the end-users according to their business requirements. The characteristic of this framework is that (i) it proposes a requirement-oriented knowledge-map and (ii) provides a methodology to match business requirements with existing components of a DIS. This methodology is supported with a case study related to the control of the manufacturing process within STMicroelectronics. A DIS is

implemented in this company to support the control of its manufacturing process with a set of indicators.

The second section in this paper presents briefly some related works. The third section presents the Topic-Map architecture of the proposed framework. The fourth section presents in detail the methodological framework with the STMicroelectronics' case study. Finally, we present in the last section the resulting Topic-Map meta-model for DIS.

2 Related work

Semantic-description techniques consist of a set of tools and methods for knowledge representation using ontology paradigm and semantic web technologies. Two main categories of technologies are used: the W3C standards such as OWL and RDF, and the ISO Semantic Web standards such as Topic Maps. These technologies are used in many disciplines to address sharing and retrieval issues.

In software engineering, the semantic description is used to enhance the retrieval of software components between application developers. Among the existing works, [6] propose an ontology-driven paradigm for component representation and retrieval. The authors propose a Component Ontology composed of five facets: Provider, Environment, Application Domain, Function and Interface. These facets can be adapted to new user requirements. They also define a retrieving algorithm based on ontology query and reasoning to enable a better syntactic and semantic matching between user requirements and component description. In [7], the authors propose a similar idea based on a description model for component ontology using OWL. This model classifies components' entities according to function entities and environment entities. Otherwise, in [8], the authors propose an RDF descriptor to capture the components' meta-data at a semantic and a syntactic level. They use an ontology model that includes source-code ontology, component ontology and a domain-specific ontology. What we can point out from these techniques is that the semantic description focuses on the functionalities and the services provided by the components. In addition, the used semantic description is flat and only provides a classification of components according to their meta-data. In the service-oriented approaches which are a related field, semantic services are proposed to add semantic to software and web services. The DAML-S technology [9] created for this purpose enables a less flat semantic description to services but this description remains also more related to a functional usage than to a business usage.

In information retrieval as well as in resources' retrieval on the web, semantic-description techniques are usually used to expose the subjects contained in the resources. Main existing approaches aim at classifying resources in grouped subjects [10] [11] [12], besides that few of them propose methodological approaches. For example, in the HyperTopic approach [13], the authors use the concept of point-of-view and entity to improve the access to resources. The point-of-view corresponds to a vision of a user or a group of users related to a requirement. The entity refers to the main subject of a resource. A collaborative-building technique is used to reference all used resources by the users. In [12], the authors propose a layered framework extended from the Topic-Map paradigm to organize knowledge with resources on four levels: the cluster level (provides the effective navigation and browsing mechanism for end-users), the Topic level (provides the main subjects contained in the resources), the knowledge-element level (provides more detailed knowledge information) and the resources level related to the physical resources. The construction of the framework is supported with a merging mechanism between these levels. In [14], the authors propose an ontology-based framework to enhance the retrieval of information resources. They propose a domain ontology to reference main concepts of a domain using RDF or OWL and with the help of domain experts. They also extract resources' meta-data which contain several concepts of the domain. The system tries then to match these meta-data with the concepts of the domain ontology using a degree of mutuality.

Thus, even if a large amount of techniques and approaches exists to support the semantic-description field, they are not really adapted to the context of DIS for two main reasons. The first one is that existing techniques provide functional descriptions or even content-oriented descriptions whereas a DIS is intended to satisfy business requirements. The second one is that most techniques do not provide a methodological support to capture and structure semantic description of components. According to these findings we propose a methodological framework to build and integrate semantic description in a DIS using the Topic Map standard. This framework is based on a requirement-oriented description for DIS components on one hand, and on a methodology to integrate the semantic description in such a system on the other hand.

3 The topic-map architecture

As introduced in section I, we use the Topic Map standard, to technically support the integration of the semantic description in a DIS. The Topic-Map concepts are used in a layered description to structure different knowledge related to both the business domain and the DIS.

3.1 The Topic Map standard

Topic Maps are an ISO semantic web standard usually used to improve information retrieval and the navigation in web resources. The key concepts of a Topic Map consist of topics, associations, occurrences and resources (figure 2).

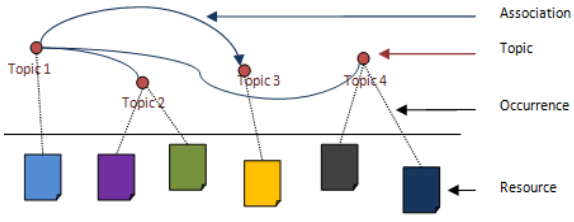


Figure 2. Core concepts of the Topic Map standard

A topic is a symbolic representation of a subject where a subject is a concept from a real world [15]. An association expresses a relationship between topics. An occurrence is what links an information resource to a topic and finally a resource is any technological support that handles information. It could be a document, a web page, software, a Database, etc.

We chose to use the Topic Map standard for several reasons. The main concepts provided by the Topic-Map Meta-Model enable to describe heterogeneous resources with high semantic abstraction and regardless of their type and their location. A Topic Map can therefore represent any subject from the real world with any desired level of granularity by typing topics and associations. Furthermore, one main characteristic of Topic Maps is that they are highly oriented towards human users comparing to the W3C standards such as OWL and RDF. These ones are more devoted to machines and interoperability between applications [16] [17]. In fact, Topic Maps are optimized for findability. The key concepts of the Topic Map paradigm enable to organize the way of navigation among resources and to effectively improve the retrieval of these resources by human users. In addition, Topic Maps are supported with many tools for topics' structuring and search.

According to our purpose, a Topic Map is used to integrate a layered semantic description in a DIS starting from business requirements and using typed topics and associations. The retrieval of such components by end-users is then driven by this description.

3.2 The Layered Topic Map for DIS

The Topic Map standard is used in our framework to provide a semantic description for the components of a DIS. We mainly use the concepts of topic and association to handle this semantic. We also use the concept of resource to represent the components that produce the indicators. At last, we consider that this semantic constitutes a part of the DIS because it normally must be integrated in such a system during its design and implementation.

Moreover, in current implemented systems, a user must know the services provided by DIS components to obtain the needed indicators. By using the semantic description, the main access entry to a DIS will be the semantic interface which results from the methodological framework. Therefore, the user can retrieve an indicator without necessarily knowing the components that provide it. As depicted in figure 3, the Topic Map constitutes the semantic interface for the access to a DIS. The Topic Map is organized in three layers: the requirement layer, the service layer and the resource layer.

- *The requirement layer:*

The main asset of the framework focuses on this layer because it gives a requirement-oriented description for a DIS. In fact a DIS provides a set of indicators and metrics to satisfy requirements in a business domain. Therefore, the semantic description must be enhanced with business requirements so to enable to better meet users' needs in terms of indicators. The topics in the requirement layer are then typed as *requirement topics*. In addition a requirement related to a business domain can be complex; hence a refinement mechanism will be used to refine a requirement until the indicators definition. Afterwards, each indicator can be provided by one or several decision-support components. The service layer with the resource layer aims at describing these components so to meet the users' requirements.

- *The service layer:*

The service layer aims at exposing the main services provided by a DIS. As introduced in the beginning of this paper, a DIS aims at processing large amount of data to produce aggregate of measures. Thus, a DIS provides a set of outputs which are the indicators sought by the end-users. The used topics in this layer are typed as service topics.

- *The resource layer:*

This layer enables the real linking to DIS components. The concepts of *occurrence* and *resource* of the Topic Map paradigm are used for this purpose. A resource represents a component from the DIS. The occurrence links a resource to a service topic with a web address or a physical address (URL, DNS, etc.). In addition an occurrence can be typed according to the types of DIS components.

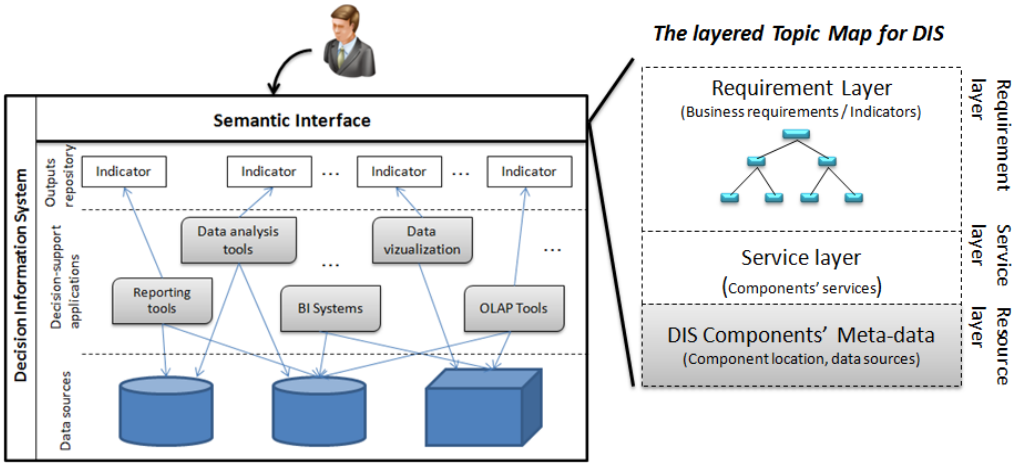


Figure 3. The DIS with the layered Topic Map

4 The methodological framework

4.1 Overview

Our methodological framework provides an environment to build and integrate semantic description in a DIS using the Topic Map standard. We use a meet-in-the-middle approach which combines a bottom-up building technique of the topic map (i.e. from DIS components) and a top-down building technique (i.e. from business requirements). The main outlines of our methodology are summarized in figure 4.

In the first step, the structure of the indicators is formalized and referenced in an *indicator ontology*. The indicators' structure is retrieved from the DIS of the company and is used afterward as support for the construction of the service layer. The construction of the requirement layer is supported with a requirement elicitation step basing on a requirement meta-model. After each requirement elicitation, the requirement model is transformed into requirement topics in the requirement layer. If same requirements are identified, some merging rules are applied. Afterwards, the requirement layer is linked to the service layer using the Topic-Maps matching rules. We note that this process aims at meeting as best as possible each requirement with an existing component.

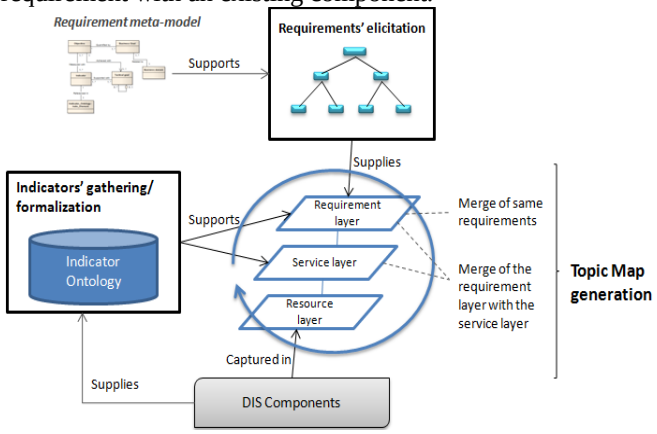


Figure 4. Outlines of the methodology

The main steps and models of the methodology are illustrated with a case study within STMicroelectronics Company. STMicroelectronics is a French-Italian company specialized in the manufacturing of electronic chips. It has implemented a DIS since many years to control the manufacturing process through a set of indicators. STMicroelectronics' product is a lot (or many lots) composed of a set of wafers where a wafer is a silicon plate used as support for the construction of chips. Each lot goes through many processing steps and is performed with production equipments. The control of the manufacturing process with indicators enables the company to ensure that at any time front-end manufacturing processes and equipments meet the requirements for products. The main outlines of the proposed methodological framework have been set with the study of the DIS implemented within STMicroelectronics. Our methodology is based on an iterative process to enable the integration of the semantic description in a consistent manner, besides that it is difficult to gather large amount of requirements in one step in a big company like STMicroelectronics. This iterative process also enables to enhance the semantic during the construction of the Topic Map.

4.2 Construction of the service layer

The challenge in integrating the semantic description in a DIS mainly resides in the heterogeneity of the existing components. A DIS integrates a large variety of data sources and applications and the produced indicators represent heterogeneous information. For this reason, in the first step of the methodology, we propose to formalize and normalize the indicator structure in an indicator ontology. The goal of this ontology is to identify all the elements that constitute an indicator in a DIS. In addition, this ontology will enable to unify the terminologies used to express a requirement in terms of indicators. As a result, the matching between a requirement and the potential DIS components that can satisfy it will be better specified. A generic ontology for the indicator structure is given in figure 5.

An indicator is composed of a set of three main elements: *Business fact*, *Dimension* and *Output-form type*.

- *Business fact:* is a measurable business concept from the business domain. It is composed of an *attribute* which represents the central business concept, a *unit of measure* and eventually a description. This description can be used to contextualize the attribute in the business domain. We propose then in this description to use a controlled vocabulary from the business domain. A process fact can be analyzed according to one or many dimensions.
- *Dimension*: represents an analysis axis for a business fact. A dimension is described by a type and a set of values related to this type.

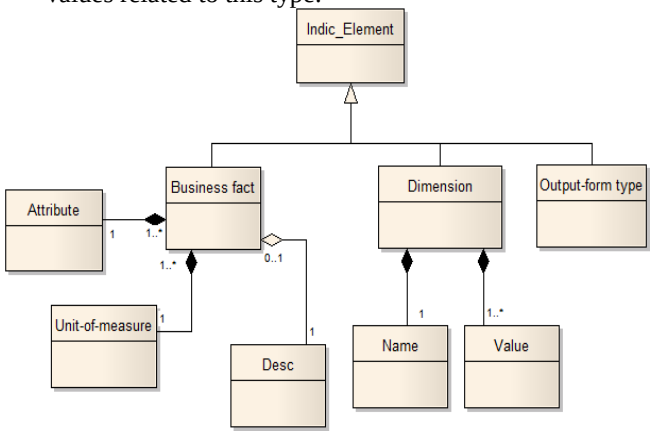


Figure 5. The generic indicator ontology

- *Output-form type:* is the type of display of data such as the type of a graph, etc.

At last, an indicator is defined by the triplet (*BusinessFact, Dimension Output-formType*) where a BusinessFact is expressed with a *unit-of-measure* and an

attribute. At last, this ontology enables to support the construction of the service layer. To that aim, we use a Topic Map to create and integrate this ontology in the service layer. We use the Wandora tool, a free software environment for Topic Maps creation and visualization. One interesting characteristic of this tool is that it enables to create layers of Topic Maps and merge them in a single Topic Map. It also enables to extract topics from different sources (XML, RDF, SQL databases, text documents, etc.). In the context of the DIS of STMicroelectronics, these techniques can be used to extract the concepts related to the indicators from the existing components. These concepts must be after structured according to the defined indicator structure. Indeed, a component takes data related to a business fact and a dimension from the business domain and produces an indicator in a type of output form, all specified by the user during the creation of the indicator. Thus, we use the service topic in the Topic Map to expose a component -referenced by its name in the service topic- and its outputs which are the produced indicators. We use three types of associations in this layer.

- *Provided by (I, C):* means that an indicator is provided by a component (the component name is used in this case).
- *Analyzed by (BF, D)*: means that a business fact is analyzed by a dimension
- *HasOutput (BF, OF)*: means that a business fact has a type of output form for display

At this step, a Topic Map for the service layer is produced and linked directly to the resources using the occurrences. Figure 6 shows an extract of the Topic Map that is being implemented within STMicroelectronics using the tree visualization in wandora.

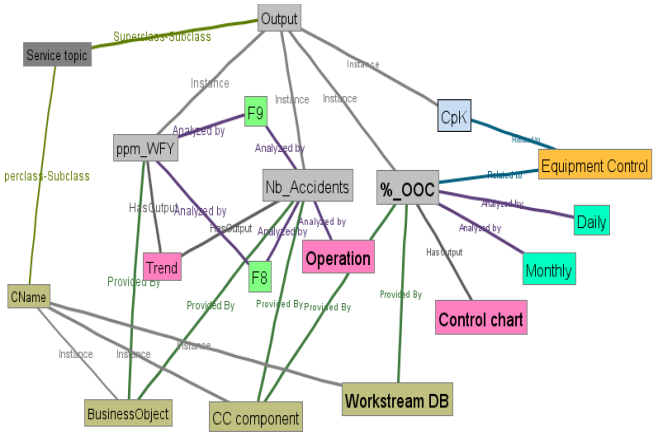


Figure 6. Extract of the Topic Map of the service layer

We have in figure 6 the triplet *(%_OOC, Daily, ControlChart*) as indicator used within STMicroelectronics for the control of equipments. An *OOC (Out of control)* is detected when the measures used to perform the products with equipments have exceeded the limits fixed by the production engineer. Thus, the concept *OOC* is specific to the function *equipment control* from the business domain. We use then this description for the attribute *OOC*. The main purpose of a description here is to enable a better referencing and retrieval of the indicators according to the business context. This description doesn't reflect the user requirement; it only gives the main function of the indicator according to the business domain. It helps the end-user to choose an indicator instead of another for some specific needs.

4.3 Construction of the requirement layer

An indicator is required by a user in a business domain to satisfy a business requirement. The nature of business requirements can differ following the business context of using DIS but the concept of *goal* that can be refined following its complexity is common to any company that has a DIS. Thus, to assist the elicitation of business requirements in the requirement layer, we propose in figure 7 a generic requirement meta-model related to a business domain. This meta-model can be adapted and enriched according to the business context of the company.

In a business domain, business goals like corporate or strategic goals are achieved by a set of quantified goals known as objectives. In the Business Motivation Model [18], an objective is attainable, time-targeted and measurable. If we take as example the control of the manufacturing process within STMicroelectronics, one of the strategic goals of the company is to *Improve the Wafer Fab Yield (WFY)*.

This goal is quantified in each STMicroelectronics' plants. For example, in the Rousset (France) plant, this goal is translated into the objective *Reducing WFY loss of x%*. An objective is achieved with a set of tactical goals.

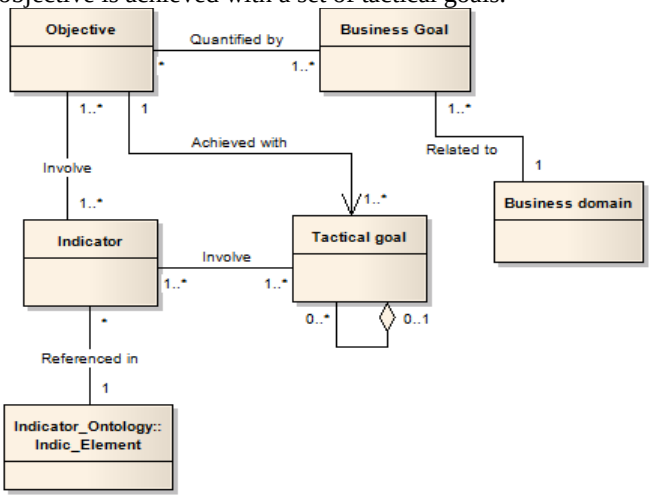


Figure 7. The generic requirement meta-model for DIS

A tactical goal defines the approach to implement and achieve an objective. This type of goal can be complex and then refined into sub-goals using a goal decomposition mechanism such as the AND/OR tree [19]. If we take again the goal of *reducing the WFY loss*, the actors involved in this process must *reduce the scraps and prevent them* by a set of control methods (a scrap is a wafer that is not compliant to the quality and the reliability requirements).

Because an objective is measurable, it has metrics based on unit-of-measures. These metrics are the i*ndicators*. Following the decomposition of the tactical goals, each goal here can involve several indicators. These indicators

represent the core of user requirements because an actor can know with the indicators if he reached business requirements. Afterwards, the indicators' definition for each low-level requirement is done using the indicators' structure defined in the indicator ontology. The requirements' elicitation is currently done within STMicroelectronics step by step by interviewing groups of business functions and analyzing their requirements. The requirement layer of the Topic Map is realized iteratively following the advancement of the requirements' elicitation. If same requirements are identified, they are then merged in the Topic Map. The associations between the requirements must also be identified to enable a consistent linking between the topics of the requirement layer. The final purpose is to obtain a single and unified model of business requirements with the required indicators for the business domain. We can see in figure 8 the example of reducing WFY loss within STMicroelectronics with the topics of the requirement layer. We identify two types of requirement topics: *goals* and *indicators*. Thus two types of binary associations are used to link these topics:

- *Require (G1,G2):*means that a goal G1 requires the goal G2 to be achieved (refers to the decomposition of goals).
- *Involve (G,I)*: means that a goal G involves an indicator I to be satisfied.

The associations *analyzed_by* and *hasOutput* are also used in this layer to specify the triplet (*BusinessFact*, *Dimension*, *Output-Form_type*) for the indicator structure.

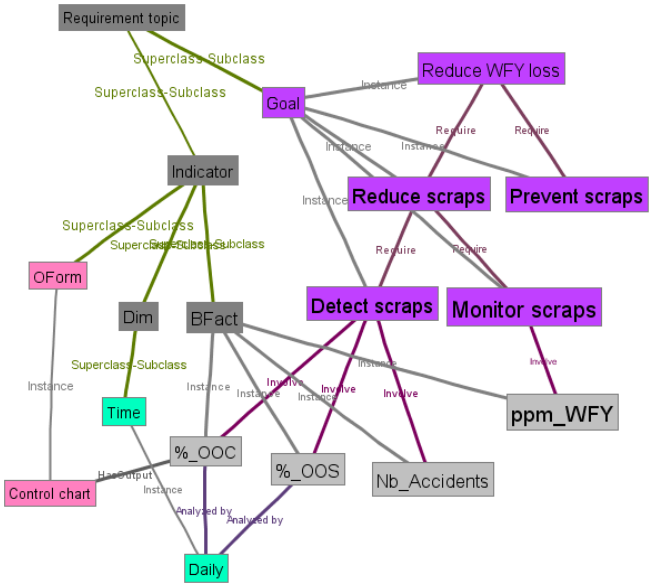


Figure 8. Extract of the Topic Map of the requirement layer

4.4 Linking the requirement layer to the service layer

The set of indicators defined in the requirement layer are matched with the outputs of the components at this step. To associate an indicator to a service topic from the service layer we use the business fact of the indicator structure.

Afterwards, a matching process is used to meet each indicator (requirement topic) with an output (service topic). One of the meaningful merging rules of the wandora tool is the syntactic matching between two topics using statistical techniques. For each indicator captured with the triplet (*BusinessFact, Dimension, Output-Form_type*) the system selects the component(s) that realize(s) the best matching with the required indicator structure. One important point in this matching is that the business fact is mandatory to satisfy at least a part of the requirement. For example, consider the indicator *(%_OOC, Daily, ControlChart*), if the *%_OOC* is given by a component but without the required dimension and output form, the system proposes anyway to link this component to the indicator. It means that the user must afterward adapt this component to his need. The main purpose of using this matching strategy is to enable reusing as much as possible existing components. At the end of this process the whole Topic Map is generated with the specified semantic. Figure 9 illustrates a part of the example *reduce WFY Loss* within STMicroelectronics. The resulting Topic Map enables an access entry to DIS components starting from any level of requirements. This Topic Map provides a semantic search to a user using the goal decomposition mechanism. This mechanism also assists the user during the definition of new requirements that involve indicators.

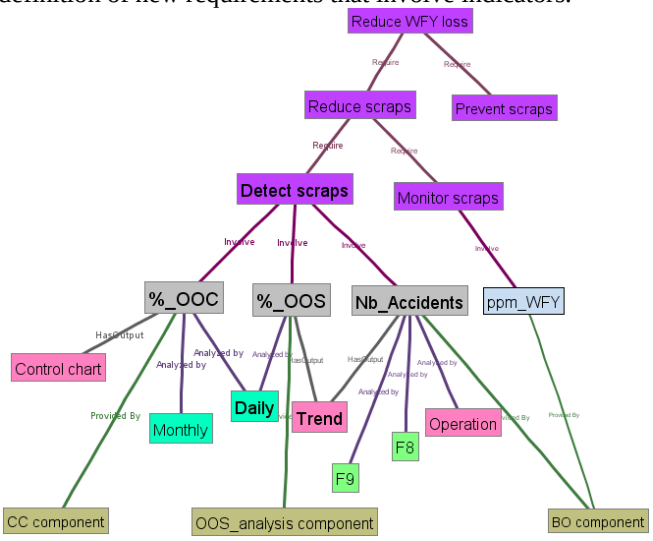


Figure 9. Extract of the resulting Topic Map

5 The resulting Topic-Map meta-model

According to the main outlines of the proposed methodological framework, new concepts have been created and used for the integration of the semantic description in the DIS. As a result, the standard meta-model of Topic Maps has been extended to include the concepts of our framework. Figure 10 presents the resulting meta-model. The standard concepts in a Topic Map are depicted with yellow color. We extended from the concept *topic* two types of topics as explained before: the *requirement topic* and the *service topic*. The associations used are extended from the association type. We note that these associations can be typed as well as needed. Finally, the proposed layered semantic description in

this framework leads us to create a new type of Topic Map. In fact the concept of layer doesn't exist in the Topic Map paradigm. Even if some Topic-Maps-based approaches try to propose layered descriptions [11] [12], these layers only refers to a group of topics, they are not implemented in practice. We propose then in this meta-model a new type of Topic Maps for DIS description (DIS TM in figure 10).

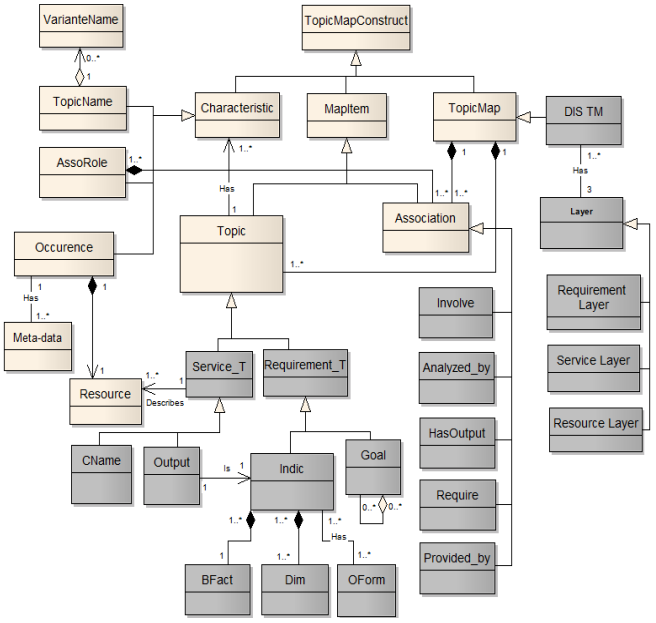


Figure 10. Meta-Model of the Topic Map for DIS

6 Conclusion

We proposed in this paper a Topic-Map-based framework to support the integration of business semantic in a DIS of a company. This business semantic aims at enhancing the retrieval of DIS components by end-users for further reuse. One main characteristic of the proposed framework is that it provides a requirement-oriented description for DIS components. In addition, the Topic Map paradigm provides a meaningful technological support for semantic sharing and retrieval of these components.

The study of business requirements within STMicroelectronics confirms us our findings and helps us to propose a suitable methodological framework to gather and integrate progressively the required semantic for a DIS. Current works try to validate the main steps of the methodology and try to study potential inconsistencies in the application of the methodology in the company. A specific search application based on the Topic Map standard is also planned to enable the effective use of the Topic Map of the framework by the end-users within STMicroelectronics.

7 References

[1] B. Rieger and A. Kleber, "Semantic Integration of Heterogeneous Information Sources : Experiences from a MSS Case 1 *," *Engineering*, no. June, pp. 89-100, 2000.

[2] L. Carneiro and A. Brayner, "X-META : A Methodology for Data Warehouse Design with Metadata Management University of Fortaleza - UNIFOR," *Production*, pp. 1-10.

[3] Y. Singh, A. Gosain, and M. Kumar, "From Early Requirements to Late Requirements Modeling for a Data Warehouse," *2009 Fifth International Joint Conference on INC, IMS and IDC*, pp. 798-804, 2009.

[4] J. Nicholas, "A Framework for Requirement Collection and Definition Process for Data Warehousing Projects Nenad Jukic Collection and Definition Process within," *Operations Management*, pp. 187-192, 2010.

[5] C. Salinesi and I. Gam, "How specific should Requirements Engineering be in the context of Decision Information Systems?," *2009 Third International Conference on Research Challenges in Information Science*, pp. 247-254, Apr. 2009.

[6] Y. Peng, C. Peng, J. Huang, and K. Huang, "An Ontology-Driven Paradigm for Component Representation and Retrieval," *2009 Ninth IEEE International Conference on Computer and Information Technology*, pp. 187-192, 2009.

[7] W. Li, Y. Guo, W. Liao, and R. Hang, "Research on Ontology Component Description Model Based on the Semantic Web," *2008 IEEE Asia-Pacific Services Computing Conference*, no. 626120, pp. 697-702, Dec. 2008.

[8] A. Alnusair and T. Zhao, "Component Search and Reuse : An Ontology-based Approach," *Knowledge Creation Diffusion Utilization*, pp. 258-261, 2010.

[9] D.-S. Coalition, "DAML-S: Semantic Markup for Web Services," 2002.

[10] C. Dichev and D. Dicheva, "Contexts as Abstraction of Grouping," *Library*.

[11] B. Sridharan, H. Deng, B. Corbitt, and B. Information, "An Ontology-Driven Topic Mapping Approach To Multi-Level Management Of E-Learning Resources An Ontology-Driven Topic Mapping Approach To Multi-Level Management Of E-Learning Resources," *Business*.

[12] H. Lu and B. Feng, "An Extended Topic Map-based Distributed Knowledge System," *Information Systems*, vol. 5, pp. 1621-1629, 2010.

[13] L. H. Zaher, J.-pierre Cahier, M. Zacklad, and I. C. Delaunay, "The Agoræ / Hypertopic approach," *Knowledge Creation Diffusion Utilization*, 2006.

[14] X. Zhang and W. Li, "Ontology-Based Semantic Retrieval System," in *2008 4th International Conference on Wireless Communications, Networking and Mobile Computing*, 2008, pp. 1-4.

[15] S. Pepper and M. Bryan, "Topic Maps," *Encyclopedia of Library and Information sciences*. 2010.

[16] S. Pepper, "Topic Maps and All That." [Online]. Available: http://topicmaps.wordpress.com/2008/05/11/topic-maps-and-the-semantic-web/.

[17] Y. Ding, M. Stollberg, and D. Fensel, "Semantic Web Languages – Strengths And Weakness," *Language*, 2001.

[18] T. B. R. Group, "The Business Motivation Model," 2010.

[19] A. V. Lamsweerde, "Goal-Oriented Requirements Engineering : A Guided Tour," in *Proceedings Fifth IEEE International Symposium on requirements Engineering*, 2001, no. August, pp. 249-263.

Online Geographic Information Systems Based Iconic Visual Query and Information Visualization Framework for Exploratory Data Analysis

Ömer M. Soysal[1,2,3], Pei Li[2], Helmut Schneider[1,2], Harisha Donepudi[2,3], Naveen K. Kondoju[2,3], Kazim Sekeroglu[2]

{[1]Department of Information Systems and Decision Sciences, [2]Highway Safety Research Group, [3]Department of Computer Science}, Louisiana State University, Baton Rouge, LA, USA

Abstract - *This paper introduces a user-friendly, iconic, visual query system that integrates Geographic Information Systems and information visualization for explanatory data analysis. This system can be used by non-technical decision makers in exploring patterns and attributes associations at five levels of hierarchy- namely, what, why, how, where and when.*

Keywords: Exploratory data analysis, geographic information systems, iconic visual query, query language, visual interface.

1 Introduction

Data analysis can be conducted in several ways, such as by a model or means of information visualization. The decision makers utilize the former for quantitative analysis and the latter for exploratory purposes. One of the aims for exploratory data analysis (EDA) is revealing relations among attributes (variables) at different levels of aggregation. Further, discovering relations among the attributes by EDA can be used for selection of 'a good set of features' to be used in a quantitative analysis. In EDA, aggregation of data plays a crucial role for searching patterns and drilling down attribute associations, as it is one of the main operations of Business Intelligence. This summary of data obtained at each hierarchical level provides a global to local picture of the problem under exploration.

The aggregation levels would be related to an event, location and time. At the event level, data exploration can be conducted to enlighten what type of events occurred, how events occurred and why certain events occurred. In addition, occurrences of events would have patterns associated with location and time; then, overall aggregation levels can be represented by five basic elements: what, how, why, where, when. A similar representation, but only with three levels of reading (what (object), where (space), when (time)), was suggested by [1]. She distinguished questions based upon these three levels as:

- when + where → what
- when + what→ where
- where + what → when

The first type of question is used to identify objects for a certain type of location and time period. As an example, "What are the traffic accidents during the rush hours on highways?" A detailed discussion of spatio-temporal data analysis topologies is given in [2].

Geographic Information Systems (GIS) is "an umbrella term for tools designed for processing, analysis, modeling, and storage" [3] of spatial data. The GIS environment provides an integrated platform for information visualization of such data. The GIS has been used widely to explore attribute relations associating them with their locations for exploring patterns in traffic accident data [4-7].

The non-visual query-based systems offer efficient searching and aggregation operations in the form of high-level programming languages such as SQL (sequential query language). This type of systems requires technical personnel, who are trained to convert a problem definition to the textual query language syntax. As a result, a user-friendly visual query system can relieve such a task for non-technical decision makers. This paper introduces a system that integrates GIS and a user-friendly iconic visual query system.

Several visual query tools, which are designed to build operation models working with GIS, have been introduced. Among those introduced are the ArcGIS model builder, AutoCAD Map 3D *Workflow Designer*, IDRISI GIS *Macro Modeler*, and ERDAS *Model Maker* were discussed in [8]. In a recent study [5], a visual programming language (VPL), which integrated with a web-based geographic application, was developed. The aim of this study was to develop a visual programming language based upon a UML model. In [9], authors developed a VPL system that makes use of "smiley icons" to access data and form queries. Similarly, [10] introduced an Iconic Visual Query Language (IVQL), which enables users to build "any combination of static, dynamic, simple and complex queries". [11] proposed a prototype visual interface to query spatial data by sketch.

2 Iconic Visual Query System

In this section, we introduce our iconic visual query system (IVQS), which is integrated with an online GIS application. Our visual query system is composed of six main modules

(icon containers), including data retrieval, filters, sorter, actions, scenario building board (SBB) and SBB operations as seen in Figure 1. First, a user needs to obtain spatial data from the map services using filters developed for the GIS application. After obtaining the data, the IVQS is used to aggregate and visualize information for exploration at hierarchical levels of five basic elements— what, how, why, where, when.

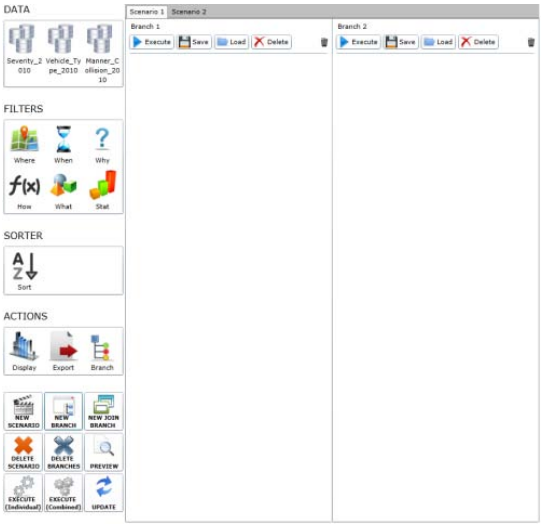


Figure 1 Main user interface

2.1 Modules

2.1.1 Data Retrieval

Data retrieval module holds the data obtained by filters located in GIS application site. Each icon in this container is linked to a map layer (category) queried in the GIS site. A thematic map category (layer) has a default query field (attribute), as well as many others in its attribute table. Each category is associated with a database table. Some categories may have a common data source. As an example, in our system, the categories 'severity', 'manner_of_collision', and 'vehicle_type' have a default field 'severity_type', 'man_col', and 'veh_type', respectively; the database table of the first two is 'Crash_table' and that of the last one is 'Vehicle_table'. Our IVQS allows multi-category and multi-data source querying.

The IVQS initializes with a SSB having two scenario panes and two branches on each by default. A user first needs to drag-and-drop a categorical data icon on to the branch from the DATA container.

2.1.2 Filters

The FILTERS container holds the five basic elements 'what, how, why, where, when' of exploring icons and 'stat' which is for statistical attributes such as 'num_inj' for number of injuries. Each iconic filter is linked to the Attribute_Filter_Type table to retrieve all corresponding attributes; as an example, the icon WHERE is linked to the table to retrieve all location related fields.

When users drag-and-drop an iconic filter, a user interface appears as seen in Figure 2. In Figure 2, after a WHERE iconic filter is dropped onto the branch, a popup window shows up; users should set a WHERE attribute, a relational operation such as 'equal', and a logical operation. All available WHERE attributes associated with the data source (e.g., Severity_2010) prior to this WHERE icon are retrieved from the database. If more than one data source exists prior to this WHERE icon, the available attributes will be the common WHERE attributes among all these data sources. When the 'Generate Filter' button is clicked, a WHERE filter string is formed if the validation procedure succeeds.

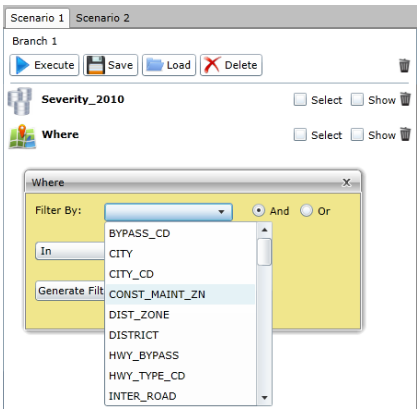


Figure 2 Adding a new filter

2.1.3 Sorting

The role of the sorting icon is to sort aggregated data obtained in the previous part of the scenario.

2.1.4 Actions

There are three action icons provided by our IVQS. The 'Display' icon is to display the aggregated query results on a chart. The 'Export' icon can be used to export the detail information of the query results into an XML file for future use. The 'Branch' icon is used to create a sub-scenario pane.

2.1.5 Scenario Building Board

Scenario Building Board (SBB) is the main module in which all the operations are executed. Users can drag-and-drop icons onto the SBB. The SBB is composed of scenario panes where each has sub-scenario branch(es) and iconic operation commands. A scenario pane can contain more than one branch; each branch holds sub-scenarios. Icons on a pane can be dragged-and-dropped individually or together. Iconic operation commands are explained in the following section.

2.1.6 Scenario Building Board Operations

Nine iconic commands for a scenario are provided in this IVQS. In addition, the branches have their own iconic commands including 'Execute', 'Load', 'Save', and 'Delete'. The scenario iconic commands are:

- 'New Scenario' command creates a new scenario on the SBB. There are two scenarios provided by default.
- 'New Branch' command creates a new branch in the current scenario. A branch supports execution,

saving and loading of the scenario on the branch, and deleting a branch.

- 'New Join Branch' command creates a new join branch. A join branch can join two query results into a new data source. This new data source can then be used for future querying.
- 'Delete Scenario' and 'Delete Branches' commands are used to delete a scenario and selected branches, respectively.
- 'Preview' command opens a new window for previewing the scenario for checking.
- 'Execute (Individual)' command executes sub-scenarios and displays the results in separate charts.
- 'Execute (Combined)' command executes sub-scenarios and displays the results in one single chart.
- 'Update' command updates the Attribute_Filter_Type table. Only administrative users can access this function.

2.2 Forming Scenario

2.2.1 Scenario Types

Our IVQS can provide several scenario types from simple to complex including *basic*, *branching*, *combining*, and *joining*; these scenarios can have single or multiple categories associated with a single or multiple database tables. The *basic* scenario type can include only one categorical data on a single branch; Figure 3 shows a sample basic scenario. In this scenario, the data from the layer 'Vehicle_Type_2010' is filtered by WHAT + WHERE + WHEN and then sorted by 'Day_Of_Wk'.

The *branching* scenario is used to add sub-scenarios under a main one. For example, a scenario may have WHAT + WHERE + WHEN at the root and another WHAT filter in the first branch and another WHEN filter in the sibling branch as seen in Figure 4. The charts in Figure 4 show the query results sorted by the attribute 'Day_Of_Wk' for each sub-scenario.

The *combining* scenario is used to compare different query results in one chart. As an example, a scenario may have WHERE + WHAT for the first data source and WHEN + WHY for the second data source: therefore the chart will display both sub-query results in one as seen in Figure 5.

The *joining* scenario is used to generate a new data set from different data sources which have a common attribute. As an example, a scenario may have one data source with WHERE + WHEN filters as well as another data source with WHEN + WHY filters; then, the joining of the sub-scenario with WHAT + WHY will display an intersection of two data obtained from two sub-queries, as seen in Figure 6. This type of scenario is also named as the 'multi-data source scenario'.

The multi-category type scenario is used if some common filters need to be applied to different data sources associated with the same data table. An example of such a scenario is seen in Figure 7. The objective of this scenario is to compare the Severity_2010 results obtained in the GIS site by applying filters SEVERITY_TYPE={FAT, INJ} and TROOP={A, B} with the Manner_Collision_2010 results obtained in the GIS site by applying filters MAN_COLL={LEFT-TURN OPPOSITE DIRECTION, REAR END, RIGHT-TURN OPPOSITE DIRECTION} And PARISH_CD={13, 61}. On the scenario pane, we dragged-and-dropped the Severity_2010 data, a WHERE icon with TROOP={A}, the Manner_Collision_2010 data, a WHEN icon with 'CR_HOUR between 6 and 18' and a WHAT icon with SEVERITY_TYPE={INJ}; as a remark, since the WHEN and WHAT icons (having 'CR_HOUR between 6 and 18' and 'SEVERITY_TYPE={INJ}') comes after the Severity_2010 data icon, the both apply to the Severity_2010 data as well. Therefore, the result on the chart can be used to compare injury crashes with 'left/right turn opposite direction or rear collision' in parish 13 and 61, versus the whole troop A with any type of collision between 6am and 4pm throughout the week. The chart can be read as 29% and 22% (4 out of 14 and 4 out of 18 injury crashes in troop A) of crashes ended up with 'right/left turn opposite direction or rear collision' between 6am and 4pm on Mondays and Tuesdays, respectively.

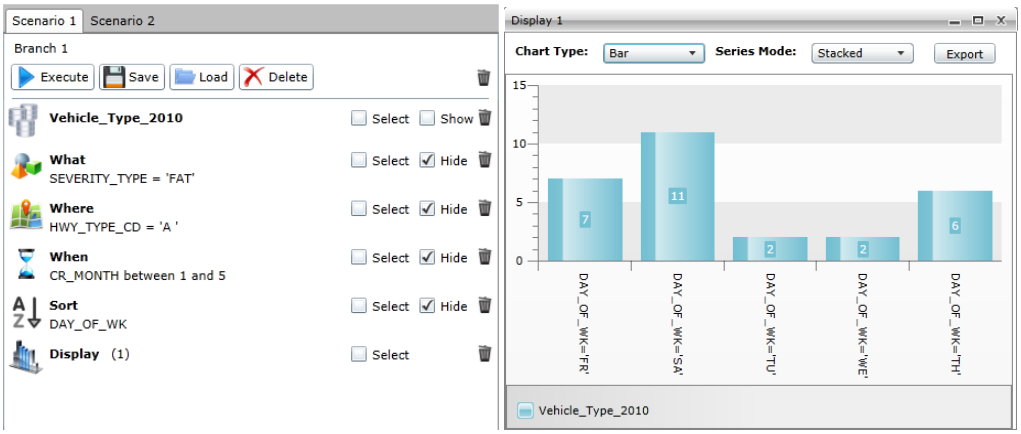


Figure 3 Left: Basic scenario query; right: Basic scenario query result

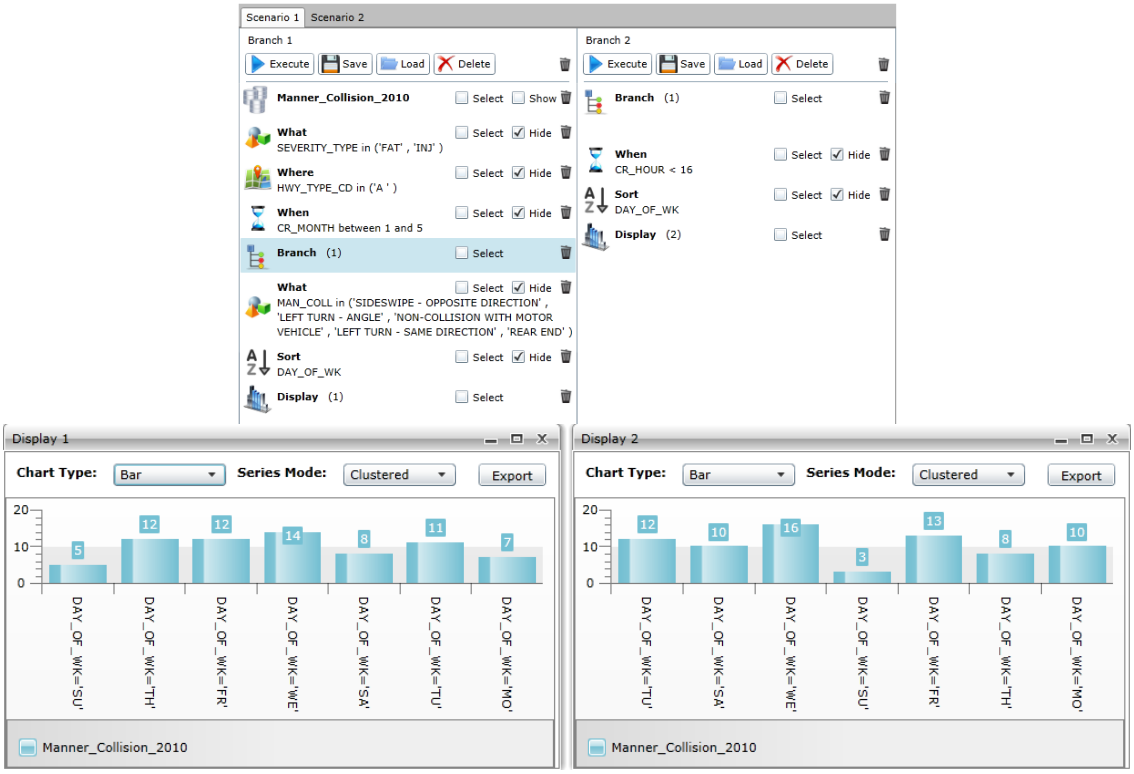


Figure 4 Top: Branch query; bottom: Branch query results

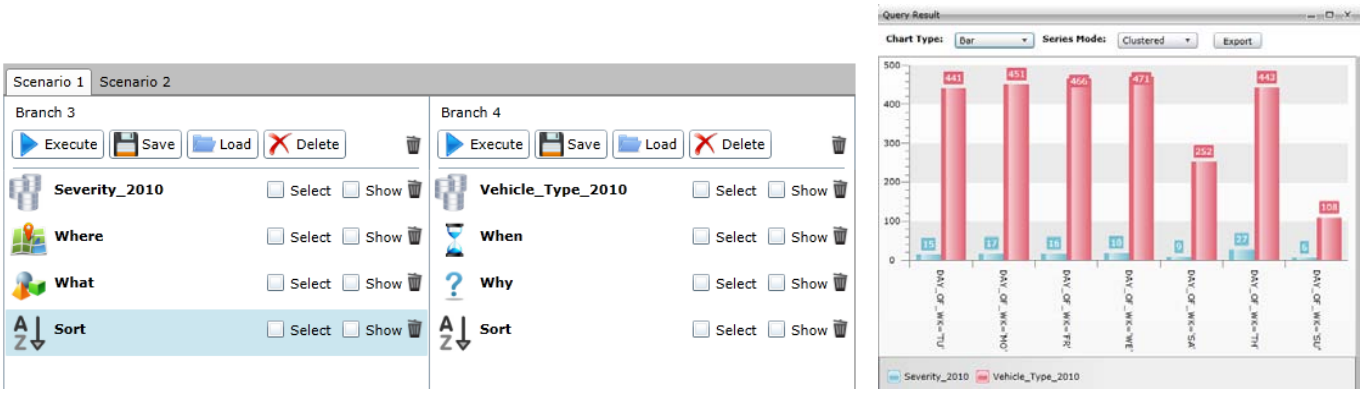


Figure 5 Two queries with different filters; bottom: Combined query results

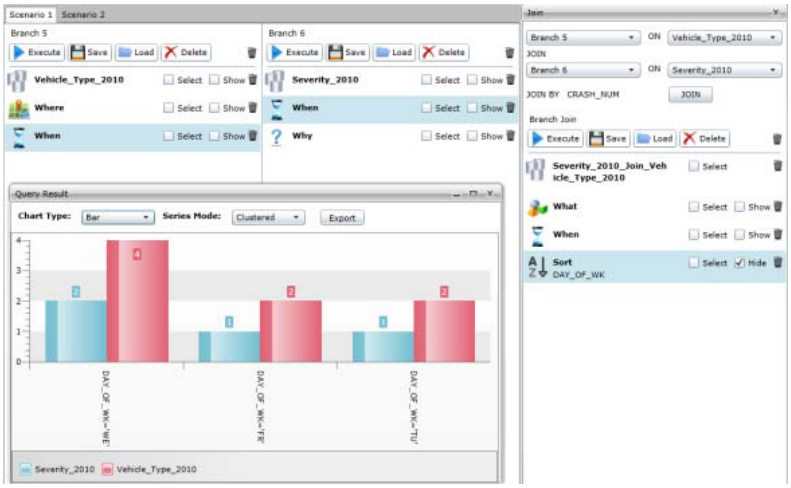


Figure 6 Join (multi-data source) query

Figure 7 Multi-category query

2.2.2 Scenario Building and Executing

An exemplary procedure of forming a scenario is as follows:

1. Drag some categorical data icons from the data retrieval module and drop to a scenario pane.
2. Drag some filters and drop into the scenario pane.
3. Drag a sorter and drop into the scenario pane.
4. Drag some actions and drop into the scenario pane.
5. Create new sub-scenario panes by clicking on the 'New Branch' button from the operations module.
6. Repeat step 1-4 for these new panes.
7. Click the 'Execute' command from the top of the each branch to execute a sub-scenario.
8. Click on the execute commands (Execute (Individual) or Execute (Combined)) in the operations module to execute all sub-scenarios of the current main scenario.

Each time an icon is placed or updated, an attribute validation procedure is conducted automatically to preserve the integrity of the query operations. The attribute validation is responsible for checking the conditions below:

1. Before the execution of a scenario, there should be some items from the data retrieval module before any items from filters, sorters, and actions modules.
2. If a branch icon is dragged from a scenario pane to another pane, it should be the first item in the second pane.
3. Items for filters and sorter modules are bound to the selected data source; if a data source is deleted, some of the filters or sorters may become invalid.
    - Invalid items are marked with pink background color; a scenario cannot execute until all the items have white background meaning validation returns no errors.

3 Conclusion and Future Work

In this paper, we have proposed an iconic visual query and information visualization system that is integrated with an online GIS site. This system can assist non-specialist decision makers for exploratory data analysis. Our system allows simple and complex query requests by means of iconic objects.

One of the limitations of our prototype system is that organization of icons needs to be sequential. We plan to enhance the scenario building board interface of our system, which would allow users to place icons in the form of a tree.

Acknowledgment: The authors of this paper would like to thank LADOTD for supporting this work.

4 References

[1] D.J. Peuquet. It's about time: a conceptual framework for the representation of temporal dynamics in geographic information systems. Annals of the Association of American Geographers, 84(3): 441–461.

[2] N. Andrienko, G. Andrienko, P. Gatalsky. Exploratory spatio-temporal visualization: an analytical review, Journal of Visual Languages and Computing, 14: 503–541.

[3] M. F. Goodchild. Geographic information systems and science: today and tomorrow, 6th International Conference on Mining Science and Technology, Procedia Earth and Planetary Science, 1: 1037–1043.

[4] S. Erdogan, I. Yilmaz, T. Baybura, M. Gullu. Geographical information systems aided traffic accident analysis system case study: city of Afyonkarahisar, Accident Analysis and Prevention, 40: 174–181.

[5] T.N. Luong, P. Etcheverry, C. Marquesuzaà, T. Nodenot. A visual programming language for designing interactions embedded in web-based geographic applications, 17th ACM International Conference on Intelligent User Interfaces (IUI 2012), pp.207-216.

[6] Ömer M. Soysal, San Chu, Helmut Schneider, Cory Hutchinson. Information Visualization Using Sub-segment Road Statistics Obtained by Coordinate Transformation in GIS Environment, 7th International Conference on Modeling, Simulation and Visualization Methods (MSV'10) in WORLDCOMP'10 -- GIS

Workshop Session on Geographic Information System Based Decision Support Systems, pp.303-306.

[7] X. Wang. Integrating GIS, simulation models, and visualization in traffic impact analysis, Computers, Environment and Urban Systems, 29(4): 471-496.

[8] Z. Dobesova. Recent Researches in Applied Informatics, Proceedings of the 2nd International Conference on Applied Informatics and Computing Theory, AICT, v11, pp.276-280.

[9] H.E. Elariss, S. Khaddaj, and R. Haraty. Towards a new visual query language for GIS, IASTED International Conference on Databases and Applications.

[10] H.E. Elariss and S. Khaddaj. Query Formulation of a Visual Query Language for Mobile GIS. Software Engineering, track: 677-056.

[11] F. Meddah, G. Hafida, B. R. Laurini. A Tool to Query GIS Based on Figurative Approach, Journal Of Computing, 3(1): 31-34.

Text Segmentation Based on Theme Analysis

Liping Yang[1], Yingwen Zhu[2], Fengxian Shen[2]
[1,2]College of Computer Science and Engineering, Sanjiang University, Nanjing 210012, China

Abstract *-In this paper, we proposed a novel-text-segmentation algorithm: first, the target text is partitioned into some blocks; then a whole-length lexical chain is constructed to analyze multiple subtopics of this text. By constructing graph, which describes blocks covering subtopics, the similar blocks, which describe same subtopic, can be classified. In this case, some segmentation points still drop inside blocks. Therefore, we take the segmentation again. Experiment results demonstrate that by analyzing theme of text, this algorithm can remove interferences, which are aroused by irrelative features, from segmentation results. By constructing graph, which describes blocks covering subtopics, it can mix similarities of adjacent and disconnected blocks together, and increases segmentation precision. The second segmentation makes segmentation results more reasonable.*

Keywords: text segmentation; lexical chain; theme of text; similarities

1 Introduction

Text segmentation is a document divided into several semantic fragments or split the unit in accordance with the linear transformation described by the document theme to form a semantic fragment of the process, enable the segmentation unit to describe the different sub-topic information [1]. Text segmentation range of applications is quite extensive, such as automatic summarization, this split of the Q & retrieval systems Chinese played a significant role [2, 3]

At present, both local fragment similarity-based text segmentation algorithm [2, 3] and global fragment similarity-based text segmentation algorithm [4, 5] are text word as clips similarity calculation basis. This method is very easy to introduce the theme has nothing to do with the document describes the noise word, increasing or decreasing the similarity between the fragment describing the different or the same sub-topic, fragment by inaccurate results. As the above algorithm does not analyze the document theme, it is likely to describe the fragments of the same sub-theme split into different segmentation unit. For the above problem, this paper presents a segmentation algorithm based on the thematic analysis of text. Remove the interference of words with the theme of information unrelated to the fragment similarity calculation algorithm by analyzing the document theme. The same time introduced to the idea of graph segmentation of the text, to make the partition into the credibility of the division for a global scope to find a connected component of the process, combined with global and local features to find the optimal value of the fragment by.

2 Class acquisition and lexical chain

Lexical chain is first proposed by Hirst in 1991, which is composed with related or similar words. Lexical chain and the structure of the text there is a correspondence between, providing important clues on the theme of the document [6]. In this paper, we select NowNet as semantic dictionary to build the semantic and lexical chain. HowNet completed by Dr. Dong is a Chinese semantic dictionary, which is defined over 1500 justice of the original, and reflects the Chinese word meaning through the meaning of the original. Each entry in HowNet [7] contain NO., The W-C and G-C, E,-C, and the DEF, of which NO concept of number, and the remaining symbols correspond to the entry word, part of speech and examples. DEF is the concept definition, expressing the entry of the semantic information consists of two parts. For example, “breath” (NO.= 084195), its DEF is {Strength: host= { human} }. The first part of the meaning is the basic meaning of the original, largely reflects the DEF; the second part is the original of the relationship between

justices, represents the relationship between the structural characteristics of DEF. HowNet the meaning of the original tree, the more similar the meaning of the original position within the meaning of the original tree the closer.

2.1 Class acquisition

HowNet represents the word meaning as the DEF, its meaning (DEF) distinction is too strict, and the basic meaning of each word DEF original largely determine the meaning of the term, at least this conclusion is established for this application. In this paper, each entry in HowNet DEF collection is divided into multiple subsets, each subset contains only the basic meaning of the original of the DEF, and a subset of as assigned to a semantic class. Each semantic class is expressed as two: the part of the semantic class corresponding to the basic meaning of the original, the other part is the relationship between justices contained in the semantic class of each DEF original union.

2.2 Lexical chain construction

Document the word and stop word filter is available after the words of the document space, through the calculation of the word space between the words reflect the degree of similarity of information will be similar or related words form a lexical chain, the following details lexical chain of the construction process:

1) Filter out HowNet the meaning is too big, too broad, abstract justice of the original, such as the "Properties", "Incident", "entity";

2) Set to be split document Doc word space for WordSet lexical chain set is L;

3) Sequential scan WordSet word set is currently being scanned for Wj, the term with t semantic classes;

4) Scanning t semantic classes of Wi, the currently j th semantic class is defined as Wij;

5) According to equation (1) Wij and L in each chain relationships, and find the chain has the greatest association with Wij, set this vocabulary chain for Lm;

6) Select value according to equation (2) calculation of Wij and Lm, such as the value of 1, then the word Wi inserted into the lexical chain Lm, and mark the Wi Lm in semantic class Wij, or the new one lexical chain includes Wi, at the same time mark Wi semantic class in the new chain for Wij;

7）If j> t, that is, Wij is the last word Wi as a righteous turn 8), otherwise the cycle steps 4)-7);

8) If the word Wi is the end of WordSet, the turn 9), otherwise the cycle steps 3)-8);

9) Calculate the weight of the L in each lexical chain, the weight of the lexical chain contains the number of words;

10) Take is greater than the average chain weight of the lexical chain as a representative of the sub-themes described in document Doc, and make a note of these lexical chain is greater than the average chain weight is a strong chain;

$$R(Wij, Lk) = \max(Sim(Wij, Lk), Cor(Wij, Lk)) \quad (1)$$

In which, $Sim(Wij, Lk)$ reflects the similarity between the information described by the Wij and vocabulary of the semantic class chain Lk; $Cor(Wij, Lk)$ reflects the correlation between the information described in Wij and Lk [8].

$$Select(Wij, Lm) = \begin{cases} 1; & \text{If R(Wij, Lm)} \geq \text{TH} \\ 0; & else \end{cases} \quad (2)$$

In which, TH is the threshold of the semantic class Wij reflect lexical chain Lm between the information, whether interrelated. (3) and (9): the word contains all the words and lexical chain maximum likelihood and maximum degree associated the word with chain. The association between the words and the lexical chain threshold is the relationship between words threshold. Found that, if the degree of association between words is over 0.6, the two words is similar to, so this set TH 0.6.

$$Sim(Wij, Lk) = \max_{p=1}^{|Lk|}(SimWord(Wij, LWkp)) \quad (3)$$

Formula (3) is the similarity between the semantic class Wij lexical chain Lk reflected. Lexical chain construction process has been marked each word in Lk corresponding semantic class in the chain, so chain Lk LWkp on behalf of the p-th word in the chain correspond to the semantic class, and Wij and Lkthe biggest similarity of all the words as in Wij and Lk similarity [9]. $|Lk|$ is the number of words in vocabulary chain Lk.

$$SimWord(Wij, LWkp) = \lambda \cdot BSim(Wij, LWkp) + (1-\lambda) \cdot RSim(Wij, LWkp) \quad (4)$$

Formula (4) is the similarity between class Wij and semantic class LWkp. This type is divided into two parts: the first part of the calculation of the degree of similarity of the two basic meaning of the original semantic class, as shown in formula (5); the second part of the calculation the degree of similarity of two semantic classes, as shown in formula (8). Parameter λ reflects the importance of the two parts. The basic meaning of the original better reflects the main message of the word, K, setting emphasis on the first part of this article Let $\lambda = 0.5$.

$$BSim(Wij, LWkp) = DL(Wij, LWkp) \cdot PS(Wij, LWkp) \quad (5)$$

Formula (5) is the similarity of the two basic meaning of the original semantic classes. The HowNet is organized in tree structure, similar to the righteousness of the original within the meaning of the original tree, the shorter the distance. We can get their similarity by calculating the two meaning of the original position within the meaning of the original tree. The first part of this formula reflects the two basic meaning of the original is in the same justice of the original tree, the second part reflects two basic meaning of the original position relationship within the meaning of the original tree.

$$DL(Wij, LWkp) = \begin{cases} 0 \\ \max(Layer(Wij), Layer(LWkp)) \end{cases} \quad (6)$$

In which, Layer (Wij) is the basic meaning of the Wij original meaning of the original tree layer.

The formula (7) calculated the relationship of the two basic meaning of the original location of the original tree of the same justice. HowNet in the meaning of the original level and the original meaning of the relationship between similarity based on at different levels of similarity between the righteousness of the original assignment [7], where interval (Wij, LWkp) refers to the semantic class Wij and LWkp to basic the meaning of the original within the meaning of the original tree-level interval.

$$PS(Wij, LWkp) = \begin{cases} 0; & if \text{ interval(Wij,LWkp)}>2 \\ 0.6; & if \text{ interval(Wij,LWkp)}=2 \\ 0.8; & if \text{ interval(Wij,LWkp)}=1 \\ 1.0; & if \text{ interval(Wij,LWkp)}=0 \end{cases} \quad (7)$$

Formula (8) calculates the meaning of the original degree of similarity of the two defined class. Relationship largely reflects the relationship between the structural characteristics of HowNet in DEF, so the formula is similar for the two defined class structure.

$$RSim(Wij, LWkp) = \frac{IS(Wij, LWkp)}{RC(Wij) + RC(LWkp)} \quad (8)$$

The size of the IS (Wij, LWkp to) the original collection of the intersection of semantic class Wij and LWkp to the relationship between justice. RC (Wij) the relationship between semantic class Wij justice the original total number.

The formula (9) calculated the correlation between the semantic class Wij vocabulary chain Lk reflect information. Similar to the similarity of word Wij and Lk related correlation as Wij and Lk.

$$Cor(Wij, Lk) = \max_{p=1}^{|Lk|}(CorWord(Wij, LWkp)) \quad (9)$$

$$CorWord(Wij, LWkp) = \frac{I(Wij, LWkp) + I(LWkp, Wij)}{RC(Wij) + RC(LWkp)} \quad (10)$$

In which, $I(Wij, LWkp)$ Wij of the righteousness of the semantic class of the original collection contains righteous class LWkp be justice of the original, such as containing a value of 1 and 0 otherwise. The basic meaning presents original justice class, and the relationship between justice on behalf of the original relationship between the characteristics of semantic class. so $I(Wij, LWkp)$ reflects the information LWkp with Wij certain.

3 Clips segmentation

This article describes the algorithm will be to split the document divided into fixed-size clips. If the end of the fragment is not re-sentence punctuation, we set the expansion of this clips with the nearest complex sentence punctuation. This method can make the end of each fragment are meaningful split points. First, the establishment of the information described by the lexical chain is a collection to

reflect the fragment for each fragment lexical chain construction method in accordance with Chapter 2 of this paper. Then, we calculated for each fragment of a collection of lexical chain intersection of a collection of strong chain with the text mode, and said that the results for the matrix A = Aij. The ith row vector of matrix A for the collection of fragments cultural strong chain and ith intersect. A behavior of a collection of strong chain as a collection of fragments, A ij the extent of information described in the document fragment covering the j-strong chain to be split. Strong chain to a certain extent, reflect the multiple sub-themes in the text, so the matrix A can be reflected to some extent, the focus of the fragment pairs theme.

Lexical chain set to be split fragments of the specified document is a collection of $BL(i)$ the size of $|BL(i)|$, the mth lexical chain of is $BL(i)_m$. The Jth-strong chain in the set text is SLj. The formula (11) is the method of calculation of the Aij.

$$A_{ij} = \sum_{m=1}^{|BL(i)|} \frac{LS(BL(i)_m, SL_j)}{|BL(i)|} \quad (11)$$

Set the lexical chain $BL(i)_m$, and cultural the strong the chain SL_j in word union for $CWSet$, its size is denoted by $|CWSet|$. The ith word CW_i the $Blockfre(i, CW_l)$ and $Articlefre(CW_l)$ for the word CW_l in the text fragments, and word frequency in the full text. If word CW_l appears in both $BL(i)_m$ and SL_j, $Same(CW_l) = 1$; otherwise, $Same(CW_l) = 0$. The similarity of the formula (12) for $BL(i)_m$ and the strong chain of lexical chain SL_j:

$$LS(BL(i)_m, SL_j) = \frac{\sum_{l=1}^{CWSet} Same(CW_l) \cdot \frac{Blockfre(i, CW_l)}{Articlefre(CW_l)}}{|CWSet|} \quad (12)$$

The first part of the formula (12) in the denominator reflects that the number of lexical chain $BL(i)_m$ and SL_j are the same word, representing the degree of similarity of the two lexical chains reflects the information. Multiplied same word CW_i in the frequency of the fragment i with the word to be split quotient of the frequency in the document, which reflects the proportion of full-text for this degree of similarity. Therefore, the formula can reflect the similarity of the lexical chain $BL(i)_m$ and SL_j.

Calculate any two fragments i, j in the matrix A corresponding row vector Ai and Aj cosine similarity, a non-directed graph reflect fragment similarity. Fragment on behalf of Figure vertex, edge (arc) represents the similarity between the two vertices (fragments), reflect this similarity of the size of the edge of the right value. Of the fragment pairs theme overlay some similarity between the vertices arc similar to the arc between the other vertices is a weak correlation, it should be based on the threshold be to remove the partition in order to reduce interference. Zhu [10] determines the distribution of word fragment internal as well as fragments segmentation threshold, but this method need to calculate every possible partition corresponding to different threshold, the calculation is very large. At the same time to be split all the words in the document is not the accuracy of the calculation contribute to the threshold value, the theme has nothing to do with the document describes the word may introduce error in the calculation of the threshold to reduce the segmentation results.

Matrix A to reflect the focus of the fragments of different sub-theme paper, it can be calculated through the matrix A fragment internal fragment between the sub-theme of the distribution of fragments within a distance between the BI and the fragment of BA, and integration of BI by linear regression and BA to determine the threshold value [11]. Distance BI can be defined as below:

$$BI = \sum_{i=1}^{b} \frac{\log_2(P_i + 1)}{b} \quad (13)$$

In which, Pi is the row vector fragment i in the matrix A corresponding non-zero number of columns. BI reflects the average value of each fragment on the focus of multiple

sub-themes in the text. Distance BA can be defined as below:

$$BA = \sum_{i=1}^{b} \frac{\|A_i - M\|^2}{b} \quad (14)$$

In which, M is the mean vector of the matrix A row vector. BA reflects the degree of dispersion of the fragments on the Documentation Center.

4 Clips internal segmentation

In above, the fragment is divided into fixed-size fragment sizes are mostly value is based on experience. So that the end of some fragments are not necessarily the true paragraphs of the split point, the real point of the split is likely to fall algorithm to determine the split point the two clips on the internal. At this point, we may fragment internal to split points in mind the split point as suspected. Fragment within the complex sentence punctuation split point as suspected, and redrawing of the fragment to those suspected of the split point. The detailed description of how to find some fall fragment of the split point is discussed as follow:

1) Set upper and lower splits of point s are $BU(s)$ and $BD(s)$, the similarity of them is $SimUD(s)$, suspected split point sets are $SegUSet(s)$ and $SegDSet(s)$;

2) The similarity sets of those suspected split point sets is $SegSimSet(s)$;

3) Scan $SegUSet(s)$, and set current scanned suspected split point with $SegU(s)_p$;

4) Set $SegU(s)_p$ as the division point, we can get two new clips: $BU(s, SegU(s)_p)$ and $BD(s, SegU(s)_p)$;

5) Calculate the intersect mode of $BU(s, SegU(s)_p)$ and $BD(s, SegU(s)_p)$, then calculate the cosine similarity of these two fragments corresponding to the intersection pattern, mark this similarity with $SimUD(s, SegU(s)_p)$, and inset it into $SegSimSet(s)$;

6) If $SegU(s)_p$ is the end of $SegUSet(s)$, turn to 7), otherwise, cycle steps 3) -6);

7) Calculate the suspected split point $SegDSet(s)$ according to steps 3) -6);

8) Set the minimum in $SegSimSet(s)$ with $SegSim(s)_{\min}$, if it is lower than $SimUD(s)$, then the suspected $SegSim(s)_{\min}$ corresponds to the split point as a new fragment of the split point.

5 Experiment and analysis

In real applications, the Chinese text of the standard split test data set, this is because the text is split generally do not separate as an independent system, but very much as part of the system for practical application. The evaluation of text segmentation algorithm is a subjective evaluation method, even for the same segmentation results of different evaluation criteria will be different evaluation results. We use search engines to retrieve 100 of 400 documents as a test corpus, as well as documentation of the test corpus manufactured calibration segmentation results.

The precision and recall rate of text segmentation algorithm similar to the accuracy of the information retrieval and recall rate, which correctly identify the split point, split point calibration and manufactured algorithm returns the split point.

$$\text{Accuracy} = \frac{\text{Correctly identify the partition point number}}{\text{Re}\,turned\ \text{point number}} \quad (15)$$

$$\text{Recall} = \frac{\text{Correctly identify the partition point number}}{\text{Correct segmentation point number in the text}} \quad (16)$$

$$F = \frac{2 \cdot P \cdot R}{P + R} \quad (17)$$

Traditional precision and recall rate is not a full and fair evaluation of text segmentation algorithm performance; the reason is that the precision and recall rate is mainly considered absolute matching results. In fact, the distant point error from the error of the correct segmentation point closer segmentation points better performance, but the precision and recall rate can not reflect this difference. In

order to overcome these shortcomings, this article also P_u evaluation methods [12] to evaluate text segmentation algorithm performance. It is defined as below:

$$P_u(ref, hyp) = \sum_{1 \le i \le j \le n} D_u(i,j)(\delta_{ref}(i,j) \overline{\Upsilon} \delta_{hyp}(i,j)) \quad (18)$$

In which, ref is the standard split mode, hyp is the segmentation model, n is the number of sentences. $(\delta_{ref}(i,j)$ and $\delta_{hyp}(i,j)$ are indicator functions. $D_u(i,j)$ is exponential distribution function of the parameter u. It is defined as below:

$$D_u(i,j) = r_u e^{-u|i-j|} \quad (19)$$

Thematic analysis of lexical chain can be seen from Table 1, our algorithm is able to remove the interference of the word has nothing to do with the topic on the segmentation results. Combination of information between the adjacent and alternate fragments of the fragment pairs theme overlay, a good predictor of a similar fragment distribution, the effect before the algorithm has been improved to some extent.

Table 1 Comparison of Segmentation results

	TextTiling	Modified Dotplot	Clips segmentation	Clips internal segmentation
			Data set 1	
P	0.39	0.34	0.48	0.51
R	0.35	0.36	0.53	0.54
F	0.36	0.39	0.52	0.53
Pu	0.59	0.61	0.72	0.78
			Data set 2	
P	0.44	0.43	0.49	0.51
R	0.33	0.39	0.42	0.51
F	0.38	0.39	0.47	0.51
Pu	0.68	0.71	0.75	0.77
			Data set 3	
P	0.42	0.36	0.47	0.52
R	0.35	0.37	0.44	0.55
F	0.39	0.36	0.47	0.52
Pu	0.70	0.69	0.71	0.78
			Data set 4	
P	0.36	0.39	0.51	0.46
R	0.39	0.42	0.43	0.54
F	0.42	0.45	0.49	0.52
Pu	0.61	0.66	0.75	0.83

The actual text on the algorithm running steps and results are described so that readers can more clearly understand the algorithm of during the execution and performance. First, the test document is divided into five sub-blocks at a fixed size. The theme focused on the degree of fragment pairs calculated by the methods described in Chapter 3, article, or intersects with the model and calculate the cosine similarity of the above fragment corresponding to the intersection mode. As shown in Table 2.

Table 2 Fragment similarity matrix

Fragment	1	2	3	4	5
1	1	0.162	0.051	0.021	0.039
2	0.162	1	0.081	0.136	0.076
3	0.051	0.081	1	0.124	0.074
4	0.021	0.136	0.124	1	0.112
5	0.039	0.076	0.074	0.112	1

The theme of the overlay graph is shown in Figure 1.

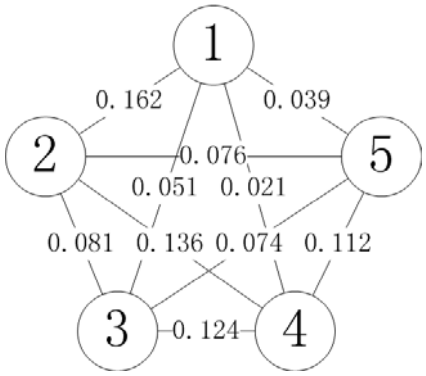


Fig.1 The theme of the overlay graph

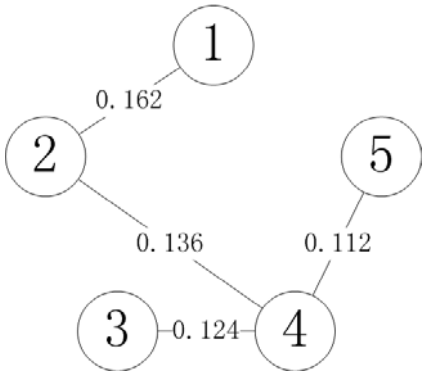


Fig.2 The manicured overlay graph

The theme of the fragment pair fusion fragment within the distance between the BI and the fragment distance BA

similarity threshold of 0.091, remove the similarity is less than the threshold edge overlay, the results in Figure 2.

In order to scan the vertices in Figure 2, first take the vertices 1, 2, we can see from Figure 2, vertices 1, 2 edge connection, 1,2 two snippet description great, it should be in the same split unit. Continue to scan vertices, take the vertex 3, that the vertices 3 and 1 and 2 describe the information of the vertices 1, 2 endless connection, we can see the clip 3 describes information and fragments, although the vertex 2 to vertex 4 edge connection. However, due to the fragment described in three sub-theme with 1,2 fragment 3 as paragraph split point, while fragments 1, 2, for a split unit. Along the vertices continue to scan top 3, 4 between the edge connecting vertices 4, 5 edge connections. We can see that fragment 3, 4 and 5 describe the information has certain relevance, and should be in the same partition unit.

6 Conclusion

Text segmentation is an important part in the natural language processing; it can be applied to many areas. However, traditional segmentation method to be split all the words in the document as the basis for segmentation calculations, thereby introducing the partition noise. This paper presents a segmentation algorithm based on the thematic analysis of text. Firstly, we take the document theme analysis by constructing the full-text lexical chain; then, we remove the irrelevant document describes the interference of the words on the segmentation results. The algorithm can also be calculated by computing the intersection pattern of the fragment of the lexical chain fragment between the similarity, and so constructed fragment pairs overlay theme. To combine adjacent and alternate fragments of information to find the optimal value of the fragment by improve the accuracy of the division. As some split point may fall on the fragment within, we take the second division to obtain a more reasonable segmentation results. Experiments show that the algorithm can make a different partition unit to describe the different sub-topic information, better segmentation results.

Reference

[1] Nikolaos Nanas, Victoria Uren, Anne de Roeck， John Domingue，Multi-topic Information Filtering with a Single User Profile，LNCS, 2004, 3025/2004, 400-409

[2] Hahn, U., The challenges of automatic summarization, Computer, 2000, Volume: 33, Issue: 11, pp: 29-36

[3] Henning Müller, Nicolas Michoux, David Bandon, Antoine Geissbuhler, A review of content-based image retrievalsystems in medical applications—clinical benefits and future directions, International Journal of Medical Informatics, 73(1), 2004, pp: 1–23

[4] K M Flaherty, D B McKay, W Kabsch, and K C Holmes，Similarity of the three-dimensional structures of actin and the ATPase fragment of a 70-kDa heat shock cognate protein，PNAS, 1991 vol. 88 no. 11 5041-5045

[5] Doug Beeferman, Adam Berger, John Lafferty, Statistical Models for Text Segmentatio, Machine Learning, Volume 34, Numbers 1-3, 1999, 177-210

[6] Young-In Song, Kyoung-Soo Han, Hae-Chang Rim, A Term Weighting Method Based on Lexical Chain for Automatic Summarization, Lecture Notes in Computer Science, 2004, Volume 2945/2004, 636-639

[7] Kok Wee Gan，Ping Wai Wong，Annotating information structures in Chinese texts using HowNet, Proceeding CLPW '00 Proceedings of the second workshop on Chinese language processing: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics, Volume 12, 2000

[8] Chan S W. Extraction of salient textual patterns: synergy between lexical cohesion and contextual coherence, IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2004, 34(2) : 205- 218.

[9] Gonenc E, Ilyas C. Using lexical chains for keyword extraction， Information Processing and Management, 2007, 43 (6)： 1705- 1714.

[10] Zhu Jingbo , Ye Na, Luo Haitao . Text segmentation modelbased on multiple discriminant analysis， Jo urnal of Software, 2007, 18(3) : 85- 94

[11] She Eryong, Wang Runsheng, Multisensor image fusion based on linear fusion model，Acta Electronica Sinica, 2005, 33(6) : 1008- 1010

[12] Beeferman D, Berger A, Lafferty J, Beeferman D, Berger A, Lafferty J, Machine Learning, Volume 34, Numbers 1-3, 177-210

Study a Web Based Intelligent Supplier Evaluation System

Jian Chen[1], Wenrong Jiang[1], and Anbao Wang[1]
[1] School of Computer and Information, Shanghai Second Polytechnic University,Shanghai ,China

Abstract - *Supplier selection is very important factor in e-manufacturing, especially a web based system which can provide instant data or result to users anywhere, anytime. The paper starts with the brief introduction of the web based internet supplier evaluation system, and then introduces methodology for selecting and standardizing elements which selected from suppliers, and finally presents the experiments results.*

Keywords: Intelligence, supplier evaluation, web based

1. Introduction

Quality, reliability, cost, flexibility, and quick responsibility are the important factors for the manufacturer to choose his suppliers in today's industry environment. An effective and correct supplier selection and evaluation tool is the key role for the manufacturer to reach the criteria of successful companies. With the development in industry, companies are facing competition globally and have the opportunities expand the business globally as well. The market requires higher quality products, best after-sale services and more fashion designed products. Furthermore, Since the World Wide Web links the each corner of the world, situation between suppliers and companies are rapidly changed. Innovation will occur in every part within the supplier selection especially for supplier selection process. The procedure of supplier selection and evaluation is to find a supplier provides best price, which can deliver in the right time with sound quality.

Since the project is to develop a web based intelligent e-supplier selection system (WBISSs) for supply chain management, which focuses on e-supplier not on traditional supplier, the biggest difference between e-supplier and traditional supplier is that e-supplier will put all of the information on the WWW but traditional supplier does not. The Figure 1 explains how the system should be.

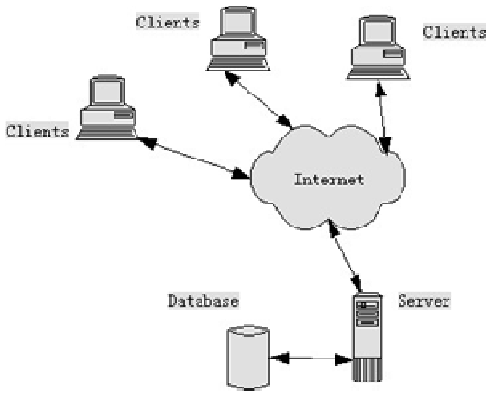


Figure 1 Internet With the E-suppliers Environment

The system should have following functions: a) System should have a stable web-server function, that clients or e-suppliers can access the information on the Web. b) System should have a database, which stores data before and after analysis. c) System is robust and secure so that it should have ability to handle heavy-duty work.

The operation of BP NN network is completed in the Matlab automatically. As a result, the WBISSs have to build a link with the Matlab to call the Matlab do the BP NN operation at the right time. Meanwhile, WBISSs should catch the output from Matlab for the further analysis. Figure 2 shows the integration of the system.

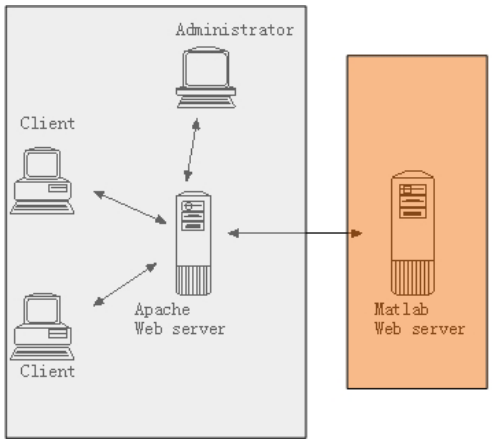


Figure2 Two Systems Integrated.

In the figure, on the left grey part is supplier selection system, which includes an Apache web server, an

Administrator, and clients. Within the supplier selection system, Apache web server deals with the requests from clients, then abstracts relevant information from World Wide Web and sent results back to clients. Administrator handles for exceptions and errors.

On the right part is the Matlab Web server, which provides intelligent ability. With Matlab Web server, Apache Web server sends results not direct back to clients but to Matlab Web server, and Matlab Web server will process these data intelligently and categories results Finally, It returns the best choice back to Apache Web server which will be then pass the best choice to the clients.

Figure 3 shows the BP model used for supplier evaluation of e-supplier system. It is a three layers system and with six input variables. There are Network Performance (NP);Price (PI); Quality (QI); Delivery (DI); Service (SI); Reputation (RI)[1]. In order to running the system successfully, the first at all is to format the input variables into the BP recognized format. It has been called input values formatting. This paper will be concentrated on the input data formatting.

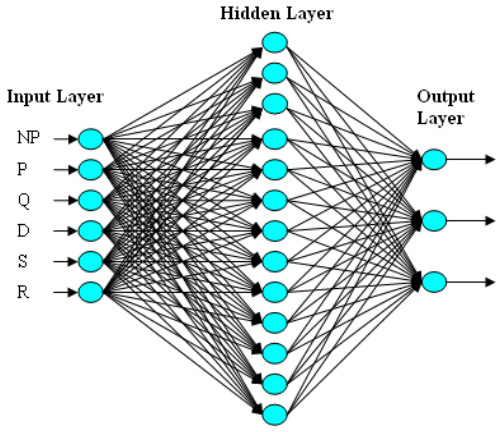


Fig.3 The BP model for the supplier Evaluation

2. Investigation the input variables of BP Neural Network

As mentioned before, there are six input variables for the BP NN model. Each variable is defined as follow:

Network Performance (NP): of each supplier can be defined according to his performance on the Internet. The criteria of the definition will be based on three factors: a.)Traffic Rank (TR): the ranking of the website; b.)Site that Link here (STLH): how many other websites that built the link to this web page; c.) Speed (SP): how fast customers can reach the website;

Price Index (PI): The price index is not only considering the price of the product, it also considering the product's quality through brand of the product. The index value will be varied according to the combination conditions: e.g. Lowest price and good brand, or lowest price and low brand, etc.

Quality Index (QI): Normally, the brands of the products imply the quality of the products. The more options imply the high quality of product. As a result, the quality index can be defined by: a) Quantity: quantity of products the supplier provides; b) Quality: What kinds of brands of products the supplier provides.

Delivery Index(DI): Delivery index is an essential issue for the supplier evaluation. Delivery index will be defined by two factors: a.) numbers of the delivery options; b.)Delivery charges of each option.

Service Index(SI): The service index is defined by: a.)How many different services the supplier provides, which includes Tel, Fax, Local Shop, MSN, Online-ticket, Free Phone, Online-form, Local shop (with e-map), support tracking, Online-chat, live support, forum, post form, online comment, e-bay account, feedback form. b.)Warranties service. c.)Return Policy service.

Reputation Index (RI): It is defined by analyzing the facts of: a.) Turnover of the e-supplier. b.) How many years the website existed.[2]

2.1 The digitizing the values of the input variables.

An e-supplier will be selected into the potential supplier pool from internet has to obtain all six input variables for the further comparing/evaluation. Table 1 shows the original collected data from internet for an e-supplier.
The data from Table 1 are very hard to comparing and input to BP NN model. It has to be convert into the numbers and the values of each variable has to be in the range of [0,1]. As a result, following calculations will be introduced respectively for each variable and the Minimum-maximum standardization formula equation 1 is used:

$$vt = \frac{v - \min_a}{\max_a - \min_a} \qquad (1)$$

Here: vt: converted value within range [0,1].
v: value of variables.
mina: The minimum value of the variable v.
maxa: The maximum value of the variable v.

According to the research, six characters [1][2] have been found to evaluate the suppliers which are the six input variables of the BP NN model.

Table 1 Typical Description of A E-supplier

Lootsale.co.uk	
Network Performance (NP)	TR:434,286
	STLH:2
	SP: VERY SLOW
PRICE (PI)	Shuttle: 14.22
	Asus:178.95
Quality (QI)	62
	Shuttle, Asus, Jetway, Lexar
Delivery (DI)	1 - 8.75
Service (SI)	Local shop, online form.
	Warranties: 12 month Return: N/A
Reputation (RI)	N/A
	N/A

2.1.1 NP: Network Performance

The first character is network performance (NP), which is very important for an e-supplier who puts all the information and transaction on the Internet. Network performance includes three sub-factors: which are Traffic Rank (TR), Site that Link here (STLH), Speed (SP).

The NP index value can be calculated by the equation Eq. 2:

$$NP = W_1 \times TR + W_2 \times STLH + \times W_3 \times SP \qquad (2)$$

Here:

TR, STLH, SP is three variables mentioned before.

is the weight for variable TR; =0.1;

is the weight for variable STLH; =0.5;

is the weight for variable SP; =0.4;

The minimum value and maximum value of TR have been defined as:

TRminimum=0;

TRmaximum=700000;

TR is used to measure the ranking of e-suppliers. Normally, rank value 1 is the best, but this comparison is different with other two variable STLH and SP. they require the highest number is the best. In order to working together with other two variables, the equation to obtain the TR's value should be slightly adjusted by Eq.3.

$$TR = \frac{TR - TR_{\min}}{TR_{\max} - TR_{\min}} \qquad (3)$$

If using the values in Table 1 as an example, the TR is 434,286, then the

TRc = 1-434286/700000 = 0.38.

The maximum and minimum values of the variable STLH is set as follow for the BP NN input. STLHmax=1479; STLHmin=1;

If the links of a home page of the supplier is more than 1479, the maximum value for this page will be fixed as 1479. Eq 1 will calculate the value of STLH. Using the same example in Table 1, the STLH is 2, so that the value of STLH is:

STLH=(2-1)/(1479-1)=0.0007.

SP can be classified into six groups: N/A, very slow speed, slow speed, normal speed, fast speed, very fast speed. Moreover, each group has its own. Here, the minimum value for SP is 0, and the maximum value for SP is 1.

Table2 Speed Description and Values

Description	Value
N/A	0.1666
Very Slow speed	0.3332
Slow speed	0.4998
Average speed	0.6664
Fast speed	0.8330
Very Fast speed	0.9996

The calculation of SP is also by Eq .1. In the same example in Table 1, SP is defined in group: very slow, so it has the value 0.3332 according to the Table 2. The SP value by Eq.1 is: SP= (0.3332-0)/ (1-0) = 0.3332。Adopted the Eq.2, the total value of NP is:

NP=0.1xTR+0.5xSTLH+0.4xSP=0.1716.

2.1.2 Calculation of the value for index (PI)

The PI index value is calculated according to the price and ranking of the brand of products. The first at all is to get the ranking of the product.

Table 3 shows the example of the ranking for the component motherboard. The score in the Table 3 will be used to calculate the Price Index (PI).

Table3 Motherboard Ranking

Motherboard ranking			
BRAND	**SCORE**	**BRAND**	**SCORE**
ABIT	**43.5**	**GIGABYTE**	**39**
ASUS	**51.5**	**MSI**	**20**
FOXCONN	**10**	**EPOX**	**32**
SOYO	**16**	**DFI**	**18.5**

In price index, there are two values are recorded:One is the cheapest price of a motherboard and the brand name, the other is the most expensive price of a motherboard and the brand name. From example, the cheapest motherboard is SHUTTLE and price is 14.22 pounds, and most expensive motherboard is Asus that price is 178.95 pounds. It is obvious that a higher ranking brand of motherboard with lower price will get reasonable higher evaluation PI value [12]. Therefore, the PI of the higher ranking of a cheaper motherboard will be obtained. The Eq.4 is used to calculate the PI value:

$$PI = \left(\frac{SCORE1}{P_{lowest}} + \frac{SCORE2}{P_{hightst}}\right) / 2 \qquad (4)$$

If Plowest > 1, the PI value will be equal to 1.

Here: Plowest is lowest price, Phighest is highest price; $SCORE_1$ and $SCORE_2$ are the value from Table 3

2.1.3 Calculation of value for Index QI

QI also contains two sections: one is number of the product ($N_{motherboard}$), and the brands of products ($B_{motherboard}$) . According to the analyzing result of the existing useful tables, the following minimum and maximum values are defined:
For value of $N_{motherboard}$: The minimum value=5; if the number of motherboard supplied by a supplier is less than 5, set minimum value=5; The maximum value=430; if the number of motherboard supplied by a supplier is larger than 430, then set maximum value=430; The value of $N_{motherboard}$ can be obtained by Eq. 1. For value of $B_{motherboard}$, the minimum and maximum values can be obtained from Table3.

The minimum=3: It means all the products are with the name SHUTTLE or ECS.

The maximum=270: It means the products include all of these brands in the Table 3.

The value of $B_{motherboard}$ can be calculated by Eq 1. During the calculation of $B_{motherboard}$, the Table 3 is required to get the score of each brand. If the brand is not appeared in the table, this brand will set the value of 1. Finally, the index value of quality can be obtained by Eq. 5:

$$QI = (N_{motherboard} + B_{motherboard}) / 2 \quad (5)$$

2.1.4 Calculating Delivery Index (DI)

Attribute Delivery (DI) contains two factors, numbers of the delivery options and the price of the delivery. In the example, the minimum numbers of delivery option is 1, and maximum is 7. The most expensive delivery charge is £53 per order and the cheapest is £0, which is free delivery[11].
The DI can be calculated by Eq.6:

$$DI = 0.35 \times (\frac{x}{53} + \frac{y}{53}) + 0.65 \times (\frac{z-1}{6}) \quad (6)$$

Here: x:Lowest delivery charge. y:Highest delivery charge. z: Numbers of options supplier provided

2.1.5 Calculating Service Index(SI)

Service (SI) contains three factors: supplier provides SI_1, and warranties SI_2, and return policies SI_3. There are totally 17 different services from the suppliers, which list in Table 4. Those different services might have different weight in Table 4.

Table 4 Services and Weights

Services	Weight	Services	Weight
E-MAIL	1	TELPHONE	1
FAX	1	LOCAL SHOP	1.5
MSN	1.5	ONLINE TICKET	1
FREE PHONE	1.5	ONLINE FORM	1
LOCAL SHOP(WITH E-MAP)	2	SUPPORT TRACKING	1
ONLINE CHAT	1.5	LIVE SUPPORT	1.5

The SI value is calculated by adding the SI_1, SI_2, and SI_3 together with the different weights as Eq.7

$$SI = 0.65 \times SI\_1 + 0.25 \times SI\_2 + 0.15 \times SI\_3 \quad (7)$$

Here, the weights for SI_1, SI_2, and SI_3 are 0.65, 0.25, and 0.15 respectively.

In Eq 7, the value of SI_1 is calculated based on Eq 1. The minimum and maximum values are defined from Table 4.

$$SI\_1_{min} = 1; \quad SI\_1_{max} = 22.$$

The SI_2 value is defined as:

$$SI\_2 = \begin{cases} 1; & warranty \\ 0 & N/A \end{cases} \quad (8)$$

The SI_3 value can be calculated by:

$$SI\_3 = \begin{cases} 1; & RP > 30 \\ RP/30; & 1 \le RP \le 30 \\ 0; & RP = N/A \end{cases} \quad (9)$$

Here: RP is the warranty days of the product.

2.1.6 Calculation of Reputation (RI)

RI has two sections: History Index (HI) of the company (how many years has been existed) :

$$HI = \begin{cases} 1; years \ge 30 \\ \frac{years}{30}; 1 \le years \le 30 \\ 0; years = N/A \end{cases} \quad (10)$$

Company's turnover (TR), There are five grades for TR: N/A, poor, average, good, excellent; (e-bay power-seller is average level), and the value of each grades are defined in Table 5.

Table 5 Five Grades for Turnover

Grade	TR number	Turnover Range
N/A	0.03	N/A
Poor	0.1	<1 million
Average	0.125	1~5 million
Good	0.25	5~20 million
Excellent	0.5	>20 million

The value of RI can be calculated by Eq. 10.

$$RI = (HI + TR) / 2 \quad (11)$$

Conclusions

PC mother board suppliers were searched on the internet and there are 10 new suppliers' data are funded as the potential suppliers.

After digitized the values of the characteristics, table 1 the digitized values of the six characteristics for e-suppliers.

Table 6: The results from the evaluation system

Company No.	Matlab returned data			Simulation results		
C01	0.997501	0.011079	0.000114389	1	0	0
C02	0.000268664	0.00000065	0.999741	0	0	1
C03	0.999841	0.00000002	0.000443405	1	0	0
C04	0.998787	0.00699871	0.000169322	1	0	0
C05	0.0011692	0.993862	0.0000000003	0	1	0
C06	0.000966617	0.994135	0.0000000003	0	1	0
C07	0.000117996	0.999991	0.0000002459	0	1	0
C08	0.999998	0.00610774	0.00000000001	1	0	0
C09	0.0000000058	0.000304284	0.999999	0	0	1
C10	0.00196454	0.949369	0.00000004985	0	1	0

References

[1] Alessandro ancarani,"Supplier evaluation in local public services: Application of a model of value for customer",Journal of Purchasing and Supply Management,Volume 15, Pages 33-42,2009

[2] Carlos Torres-Fuchslocher,"Understanding the development of technology-intensive suppliers in resource-based developing economies" Research Policy, Volume 39, Pages 268-277,2010

[3] Chung-Chi Hsieh, Yu-Te Liu,"Quality investment and inspection policy in a supplier–manufacturer supply chain",European Journal of Operational Research, Volume 202, pp. 717-729, 2010

[4] Jian Chen “Internet based Intelligent Supplier Selection and Evaluation System”, PhD Thesis, University of Derby, U.K., 2007

[5] Jian, Chen, M.H.Wu “Developing an Internet Based Intelligent Supplier Selection System”, Applied Mechanics and Materials (Volumes 10 - 12), p 58-61，2008

[6] Humphreys, Paul, Huang, George, Cadden, Trevor “ A web-based supplier evaluation tool for the product development process”, Industrial Management & Data System, Vol.195 No 2, pp.147-163, 2005

[7] Khurrum, S. B and Faizul, H. (2002), “Supplier selection problem: a comparison of the total cost of ownership and analytic hierarchy process approaches”, Supply Chain Management: An International Journal, Vol.7 No.3, pp.126-135

[8] Seong-Jong Joo, George H. Messer Jr, Ronald Bradshaw,"The performance evaluation of existing suppliers using data envelopment analysis",International Journal of Services and Operations Management,Volume 5, Pages:429 - 443,2009

[9] William Ho, Xiaowei Xu, Prasanta K. Dey,"Multi-criteria decision making approaches for supplier evaluation and selection: A literature review",European Journal of Operational Research, Volume 202, Pages 16-24,2010

[10] I. Chamodrakas, D. Batis, D. Martakos,"Supplier selection in electronic marketplaces using satisficing and fuzzy AHP", Expert Systems with Applications, Volume 37, Pages 490-498,2010

[11] Paul W.Th. Ghijsen, Janjaap Semeijn, Saskia Ernstson,"Supplier satisfaction and commitment: The role of influence strategies and supplier development",Journal of Purchasing and Supply Management, Volume 16, , Pages 17-26,2010

SESSION

MINING OF DATA RICH SOURCES

Chair(s)

Prof. Ray Hashemi

STUDY OF FEATURE SELECTION ALGORITHMS FOR TEXT-CATEGORIZATION

Kandarp Dave
University of Nevada, Las Vegas
4505 S. Maryland Pkwy.
Las Vegas, NV 89154
davek@unlv.nevada.edu

Kazem Taghva
University of Nevada, Las Vegas
4505 S. Maryland Pkwy.
Las Vegas, NV 89154
Kazem.Taghva@unlv.edu

Abstract

This paper will discuss feature selection algorithms for text-categorization. Feature selection algorithms are very important, as they can make-or-break a categorization engine. The feature selection algorithms that will be discussed in this paper are *Document Frequency*, *Information Gain*, *Chi Squared*, *Mutual Information*, *NGL (Ng-Goh-Low) coefficient*, and *GSS (Galavotti-Sebastiani-Simi) coefficient*. The general idea of any feature selection algorithm is to determine the importance of words using some measure that can *keep* informative words, and *remove* non-informative words, which can then help the text-categorization engine categorize a document, *D*, into some category, *C*. These feature selection methods are explained, implemented, and are provided results for in this paper. This paper also discusses how we gathered and constructed training and testing data, along with the setup and storage techniques we used.

1 Introduction

With the growth of online information, text-categorization has become a very important technology to categorize a large number of documents. The idea of text-categorization, or text-classification, is to categorize textual data into one or more predefined categories [1, 2, 3, 4]. Text-categorization is a "supervised technique that uses labeled training data to learn the classification system and then automatically classifies the remaining text using the learned system" [4].

Feature selection is an important part of text-categorization, and much research has been done on various feature selection algorithms. The idea of feature selection, in simple words, is to determine the importance of words using some measure that can *keep* informative words, and *remove* non-informative words, which can then help the text-categorization engine.

The feature selection methods that are studied, implemented, and provided results for in this paper, are the following: *Document Frequency*, *Information Gain*, *Mutual Information*, *Chi Square*, *NGL (Ng-Goh-Low) Coefficient*, and *GSS (Galavotti-Sebastiani-Simi) Coefficient*. These algorithms have been studied before, mainly on Reuters and Newsgroup input data. We did not use Reuters or Newsgroup data, instead, for our needs, we built custom (mixed) data ourselves.

This paper is organized as follows. Section 2 explains how we collected and tagged documents, along with the setup we used. It also explains how we created the BOW, or bag of words. Section 3 talks about how documents are counted. Section 4 explains why feature selection algorithms need to be used, and gives explanations about each algorithm. Section 5 shows the results we achieved. The paper is concluded in section 7.

2 Data Collection

2.1 Setup

We chose to use PHP as an application layer, as it takes care of minute details of implementation by providing high-level interfaces, and objects such as arrays that can be associative. For storage, we used a MySQL database. We chose not to use plain text files due to the amount of data we knew we had to deal with. MySQL and PHP work very well together as PHP has a built in connector that can easily access MySQL database. Also, it is much easier to insert,

update, and retrieve thousands of rows of data into and out of MySQL, especially with transaction ability, and SQL. Another big positive point with using MySQL is that we can easily index data, which makes searching of the indexed columns really fast. MySQL turned out to be the best choice to use than to store data in plain text files.

2.2 Categories

We did not want to have categories that were all completely separate from each other. Categories "Technology" and "Food" are considered very separate from each other. Categories such as "JavaScript" and "PHP" are considered very close to each other. Meaning, we wanted have a mix of categories where some categories would be very close to each other and some other categories that would be very separate from each other. The categories we worked with and what they are about are:

- "Chinese" - Food.
- "Indian" - Food.
- "Italian" - Food.
- "India" - General news.
- "Apple" - Technology news.
- "Google" - Technology news.
- "Facebook" - Technology news.
- "PHP" - Technology.
- "JavaScript" - Technology.

2.3 Collecting Data and Tagging Documents

Data was manually gathered from various online sources (websites) with the help of some utilities, made by us, that could help us gather data faster. 1,010 training documents and 338 testing documents were manually collected and tagged with appropriate categories.

To gather training data, 1,010 documents were determined that were appropriate enough to be categorized under the categories listed above. First the URLs for each of these these training documents was gathered. We decided to manually go through all 1,010 training documents, and gather text-content from each document that best represented that document. For the 338 test documents, we manually went through each to gather text-content that we could test the feature selection algorithms on. Using a utility program, we also collected the URL, the meta information, and the title of each document for both training and testing purposes.

Training documents were tagged for training, and test documents were tagged to check if our results were correct, and how much. Documents were tagged using the same utility mentioned above.

The utility updated the URL, the title, and the meta information in the storage, and it also linked, or tagged, each document with appropriate category.

2.4 Database Setup

We used a MySQL database, and set up appropriate tables. These tables, along with the column names and their descriptions, are explained in each section below.

2.4.1 Table "category"

The "category" table contains categories.
The columns are:

- PK_CATEGORY - Primary key.
- NAME - Name of the category.

2.4.2 Table "document"

The "document" table contains all training documents.
The columns are:

- PK_DOCUMENT - Primay key.
- URL - The document's URL.
- TITLE - The title of the webpage.
- META - Meta of the webpage. We use the value of *content* attribute of the meta tag whose *name* attribute has the value "description".
- DOCUMENT_DETAIL - This column contains the actual text-content extracted manually.

2.4.3 Table "document_category"

The "document_category" table is a linking table that links training documents to categories.
The columns are:

- PK_DOCUMENT_CATEGORY - Primary key.
- CATEGORY_ID - ID of the category to which this training document belongs.
- DOCUMENT_ID - The ID of the document.

2.4.4 Table "document_test"

The "document_test" table contains all test documents.
The columns are:

- PK_DOCUMENT_TEST - Primary key.
- URL - The document's URL.
- TITLE - The title of the webpage.

- META - Meta of the webpage, extracted the same way as it is done with the training data.
- DOCUMENT_DETAIL - This column contains the actual text-content extracted manually.

2.4.5 Table "document_category_test"

The "document_category_test" is a linking table that links testing documents to categories. When the text-categorization algorithm is run, the results from the algorithm are compared to this table. This helps determine how many documents are *true positives*, *false positives*, and *false negatives*.
The columns are:

- PK_DOCUMENT_CATEGORY_TEST - Primary key.
- CATEGORY_ID - ID of the category to which this test document belongs.
- DOCUMENT_ID - The ID of the document.

2.4.6 Table "bow"

The "bow" table contains ungrouped bag of words. Each cleaned and stemmed word from all the training documents is inserted in this table. This table is very helpful in determining document counts.
The columns are:

- PK_BOW - Primary key.
- WORD - This column contains a non-unique list of words.
- DOCUMENT_ID - Document ID for the given word.
- CATEGORY_ID - ID of the category in which this document belongs.

2.4.7 Table "bow_feature_selection"

The "bow_feature_selection" contains a grouped list of bag of words for training documents. The columns are:

- PK_BOW_FEATURE_SELECTION - Primary key.
- WORD - A list of words gathered and grouped from the "bow" table.
- CATEGORY_ID - ID of the category in which this word belongs.
- A - Number of documents in CATEGORY_ID, C, containing WORD, w.
- B - Number of documents *not* in CATEGORY_ID, C, containing WORD, w.
- C - Number of documents in CATEGORY_ID, C, *not* containing WORD, w.
- D - Number of documents *not* in CATEGORY_ID, C, *not* containing WORD, w.
- INFORMATION_GAIN - Calculated Information Gain values.
- CHI_SQUARE - Calculated Chi Square values.
- MUTUAL_INFORMATION - Calculated Mutual Information values.
- NGL - Calculated Ng-Goh-Low coefficient values.
- GSS - Calculated Galavotti-Sebastiani-Simi coefficient values.

2.4.8 Table "bow_test"

The "bow_test" table contains a raw list of bag of words for test documents. In actual testing, this table would *not* be created. The only reason for creating this table is so that we can run multiple tests on multiple feature selection algorithms easily. The columns are:

- PK_BOW_TEST - Primary key.
- DOCUMENT_ID_TEST - ID of the test document in which this word belongs.
- WORD - A cleaned word, which will help in testing of the text-categorization engine and the feature selection algorithms.

2.5 Creating Bag of Words

We created the bag of words, or BOW, for both the training dataset and the test dataset. The BOW of the training dataset is then used to train the categorization engine. The BOW of the test dataset is optional, meaning it is not required to create it, but since we wanted to run many tests, we created the BOW of the test dataset.

To create BOW, we used the b8 lexer. b8 is a Naive Bayesian Spam filter library written by Tobias Leupold. We extracted the lexer from the library, and added the stopword removal capability. Stopwords, or "overly common words", are not helpful in categorization, as they cannot differentiate between categories [5]. Words such as *the*, *a*, *of*, *is*, *at*, *on* and many more are considered stopwords.

We also added the ability to stem words using the Porter Stemmer. "The Porter stemming algorithm (or 'Porter stemmer') is a process for removing the commoner morphological and inflexional endings from words in English. Its main use is as part of a term normalisation process that is usually done when setting up Information Retrieval systems" [6]. George Forman says that "the common practice of stemming or lemmatizing - merging various word forms such as plurals and verb conjugations into one distinct term -

also reduces the number of features to be considered." [5].

We also modified the b8 lexer to replace each of the following symbols in a document with a space letter.

```
~ ‘ ! @ # $ % ^ & * ( ) - _ = +

[ { ] } \ | ’ " ; : , < . > / ?
```

The reason for replacing all symbols with a space is that a phrase such as "`MySQL's awesome`" can be converted into "`MySQL s awesome`". Letter 's' by itself is a stopword, and would be removed, leaving a descriptive phrase "`MySQL awesome`". The result will be two tokens "`MySQL`" and "`awesome`".

The b8 lexer also associates with each token, the document ID and the category ID. All tokens along with their category ID and document ID are stored in the "bow" table.

From this "bow" table, we create "bow_feature_selection" table, where rows are grouped by word and category ID. We simply run the following query to create "bow_feature_selection".

Listing 1: Query to Insert Into "bow_feature_selection" Table

```
insert into bow_feature_selection
      ( word , category_id )
   select word, category_id
     from bow
     group by word, category_id
```

The BOW for test data was created exactly the same way as for training data using the same b8 lexer.

The "bow" table had 88,230 rows and the "bow_feature_selection" table had 20,419 rows.

3 Counting Documents

Document counting is an important task in feature selection algorithms. Document counts can help determine how important or not-important a feature is. Next we explain the A, B, C, D values and how they are calculated.

3.1 Explanation of the A, B, C, D Values

- A - the number of documents in category, C, containing word/token, t.
- B - the number of documents not in category, C, containing word/token, t.
- C - the number of document in category, C, not containing word/token, t.
- D - the number of documents not in category, C, not containing word/token, t.

[4]

3.2 Calculating the A, B, C, D Values

To calculate the A, B, C, D values, we used the explanation given above to run appropriate queries. Query to calculate A is given below. Others are similar, replace "=" in the query with "<>" when doing *not in category* or *not containing word*.

Listing 2: Calculate A

```
select count(0) as A from (
   select distinct document_id
   from bow
   where category_id='".$categoryID."'
   and word='" . $word . "'
) as tbl
```

Calculating these A, B, C, D values is the most important task in feature selection. On a machine with 2.2 GHz Intel Core i7 processor, 4GB of memory and 468GB hard disk space available, it took us over 30 minutes to calculate the A, B, C, D values for all features.

4 Feature Selection Algorithms

In this section, we describe and implement six different feature selection algorithms.

4.1 Why Use Feature Selection?

Feature selection is "selecting a subset of the features available for describing the data" [7], or in other words it is a method to reduce "the dimensionality of the dataset by removing features that are considered irrelevant for the classification" [8].

There are many benefits to using feature selection, as listed below:

1. Simplifying or speeding up computations with only little loss in classification quality. [7]
2. Reduce dimensionality of feature space and improve the efficiency, performance gain, and precision of the classifier. [2, 9, 8]
3. Improve classification effectiveness, computational efficiency, and accuracy. [9, 1]
4. Remove non-informative and noisy features and reduce the feature space to a manageable size. [10]

5. Keep computational requirements and dataset size small, especially for those text-categorization algorithms that do not scale with the feature set size. [8]

4.2 Document Frequency

Document frequency is a very simple feature selection method. Document frequency for a term can be found by counting the number of documents in which a term/feature occurs. [1, 9, 3]. We have already calculated document frequency as the A column in "bow_feature_selection" table.

4.3 Information Gain

According to Mukras *et al.*, "the idea behind IG is to select features that reveal the most information about the classes" [11].

Information gain values were calculated as follows:

$$IG(t,c) = \sum_{c \in \{c_i, \bar{c_i}\}} \sum_{t \in \{t_k, \bar{t_k}\}} P(t,c) \cdot log \frac{P(t,c)}{P(t)P(c)}$$

[8]

4.4 Mutual Information

Mutual information method assumes that the "term with higher category ratio is more effective for classification" [1].

Mutual information can be calculated as follows using our already calculated A, B, C, D values:

$$MI = \log \frac{A \times N}{(A+C)(A+B)}$$

[1]

Here, A is the number of documents that contain the term, t, and also belong to category, c. B is the number of documents that contain the term, t, but do not belong to category, c. C is the number of documents that do not contain the term, t, but belong to category, c. N is the number of training documents. [1, 3].

4.5 Chi Square

Chi square measures the lack of independence between a term, t, and the category, c [9, 3].

Chi square, χ^2, can be calculated as follows, again, using our previously calculated A, B, C, D values:

$$\chi^2 = \frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)}$$

[4]

4.6 NGL (Ng-Goh-Low) Coefficient

NGL Correlation Coefficient (CC) is a variant of χ^2 metric. Uchyigit and Ma tell us that, "the NGL coefficient is reported to have better performance than χ^2" [12] They say so, because NGL "selects words that correlate with c (i.e. are positive) and does not select those words which correlate with $\bar{c}$, unlike the χ^2 statistic" [12]. The NGL CC value can be computed as follows:

$$NGL = \frac{\sqrt{N} \cdot (AD-CB)}{\sqrt{(A+C)(B+D)(A+B)(C+D)}}$$

[8]

4.7 GSS (Galavotti-Sebastiani-Simi) Coefficient

Galavotti-Sebastiani-Simi propose a *simplified* χ^2 statistic. They remove the $\sqrt{N}$ factor, and the denominator completely. They describe the $\sqrt{N}$ factor as being unnecessary. They also remove the denominator, $\sqrt{(A+C)(B+D)(A+B)(C+D)}$, by giving the reason that the denominator gives high Correlation Coefficient score to rare words, and rare categories [12]. The GSS CC value can be computed as follows:

$$GSS = AD - CB$$

Next we provide the results we achieved using these feature selection algorithms.

5 Results

We calculated *precision*, *recall*, and *F1* values to determine how accurately documents were categorized when using different feature selection methods. We can calculate *recall* and *precision* as follows:

$$recall = \frac{TP}{TP+FN}$$

$$precision = \frac{TP}{TP+FP}$$

Here, TP is the number of true-positives, FP is the number of false-positives, and FN is the number of false-negatives.

A true-positive is when a human and the categorization algorithm both agree on the result of the categorization.

A false-positive is when a human and the categorizer disagree on the result of the categorization, but by a small degree. For example, a human knows

a document belongs to "Indian" food category, but the categorizer classifies it under "Italian" food category. "Indian" food and "Italian" food are related to "food", and this is considered a false-positive.

A false-negative is when a human and the categorizer completely disagree on the result. For example, a human knows that a document belongs to "Google" category, but categorizer classifies it under "Chinese" food category, which is completely wrong, and this is a case of false-negative.

F1 is the harmonic mean of precision and recall and is calculated as follows:

$$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

5.1 Using All Features

We tested the feature selection algorithms on 338 test documents with all 20,419 features selected. We wanted to get as high recall values as possible. Results are given below (PR=Precision, RE=Recall):

Algo.	TP	FP	FN	PR	RE	F1
DF	232	106	0	0.686	1	0.814
IG	276	56	6	0.831	0.979	0.899
χ^2	313	25	0	0.926	1	0.962
MI	299	38	1	0.887	0.997	0.939
MI*	304	34	0	0.899	1	0.947
NGL	301	37	0	0.891	1	0.942
NGL*	309	29	0	0.914	1	0.955
GSS	253	85	0	0.749	1	0.856

Algorithms marked (*) are optimal runs. Meaning, not all features were used, but rather only a small set of features from 20,419 total features. Next we select only important features, and give the results.

5.2 Using Selected Features

Recall that we have a total of 20,419 features, and 338 test documents.

5.2.1 Document Frequency

# of Features Used	# of Features Removed	TP
20,419	0	232
1,714	18,705	218
729	19,690	201
412	20,007	182
261	20,158	179
175	20,244	164
127	20,292	135

Document frequency is not a very useful feature selection algorithm. But even for a simple feature selection algorithm, such as DF, after removing 18,705 features, we still got *TP* of 218. Meaning even after removing 18,705 features, we lost only 14 *TP*.

5.2.2 Information Gain

# of Features Used	# of Features Removed	TP
20,419	0	276
6,184	14,235	276

Notice that we still have the same number of TP after removing 14,235 features.

5.2.3 Mutual Information

# of Features Used	# of Features Removed	TP
20,419	0	299
19,770	649	299
18,849	1,570	300
13,509	6,910	304
7,374	13,045	286
4,921	15,498	188
955	19,464	56

Using features with positive *mutual information* value yielded better results. The point here to notice is that, even after removing 6,910 feature, we got more *TP* than when we used all features.

5.2.4 χ^2

# of Features Used	# of Features Removed	TP
20,419	0	313
4,916	15,503	313
2,039	18,380	311
1,270	19,149	308
635	19,784	306
310	20,109	305
264	20,155	304
186	20,233	303
138	20,281	306
123	20,296	300
106	20,313	302
95	20,324	300
80	20,339	302
65	20,354	303
61	20,358	302
52	20,367	283

Again, χ^2 feature selection algorithm performed very well even with less amount of features. Even

when we removed 15,503 features, and used only 4,916 features, results (*TP*) did not change. The number of *TP* stay above 300 even when we use only 61 features.

5.2.5 NGL Coefficient

# of Features Used	# of Features Removed	TP
20,419	0	301
20,223	196	305
17,061	3,358	307
10,737	9,682	306
6,078	14,341	309
2,572	17,847	305
1,540	18,879	304
995	19,424	296
218	20,201	298
157	20,262	304
91	20,328	300
70	20,349	304
50	20,369	282
36	20,383	231
26	20,393	213
20	20,399	184

As can be easily seen from these results, NGL performs almost as well as the χ^2 algorithm.

5.2.6 GSS Coefficient

# of Features Used	# of Features Removed	TP
20,419	0	253
20,345	74	252
7,141	13,278	250
393	20,026	220
132	20,287	192
50	20,369	135
5	20,414	73

For GSS, even after removing 13,278, we lost only 3 *TP*.

5.3 Explanation of the Results

Kotcz, Prabakarmurthi, and Kalita have also shown that, "by reducing the feature space, the accuracy of a classification method can be increased and, even when only very few of the original features are kept, good accuracy can maintained" [10]. Our results agree. We have shown that by *keeping* only the informative features, and *removing* all other non-informative features, we can either improve results, *TP*, or can get same results by reducing the feature set by a very large degree. As can be seen from the results that even after removing many features, we were still able to get *TP* that were close to the optimal.

Conducting one sample T-Test on precision values, with *hypothetical mean* of 0.9483, we got P value of 0.0362, and this indicates that the difference is statistically significant.

6 Conclusion and Future Work

Text categorization is very important, but we believe, the problem of feature selection is as much, or more important than text-categorization. In this paper, we discussed many important topics ranging from collecting data, to organizing data and ultimately using the organized data to efficiently conduct tests using the feature selection algorithms.

We showed how we used a MySQL database to efficiently store our collection of documents. We then described what setup was used for the database, and described the structure of each table in the database. Then we explained the details of the *b8* lexer, and described how we used the *b8* lexer to create bag of words, or BOW, to help us train and test data. Next, we explained what the A, B, C, D values were, and how we calculated each of those values using the BOW. Then, we gave explanation on the following feature selection methods: *Document Frequency*, *Information Gain*, *Mutual Information*, *Chi Square*, *NGL (Ng-Goh-Low) Coefficient*, and *GSS (Galavotti-Sebastiani-Simi) Coefficient*. Next, we gave the results. We then went on to show that even after removing features, and in some cases more than 90%, we were still able to maintain over 99% of *TP* in our results. This study has shown how powerful feature selection algorithms can be.

An area that George Forman and we believe that could be further researched is that of "hierarchical categories". Forman says the following:

> Hierarchy is among the most powerful of organizing abstractions. Hierarchical classification includes a variety of tasks where the goal is to classify items into a set of classes that are arranged into a tree or directed acyclic graph.
>
> [13]

We strongly believe, conducting further research on "hierarchical categories" would be very helpful, as it can ultimately help categorize documents on the web.

References

[1] S. Li, R. Xia, C. Zong, and C.-R. Huang, "A framework of feature selection methods for text categorization.," in *ACL/AFNLP'09*, pp. 692–700, 2009.

[2] W. Shang, H. Huang, H. Zhu, Y. Lin, Y. Qu, and Z. Wang, "A novel feature selection algorithm for text categorization," *Expert Systems with Applications*, vol. 33, no. 1, pp. 1–5, 2007.

[3] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," pp. 412–420, Morgan Kaufmann Publishers, 1997.

[4] M. Z. Fadi Thabtah, Mohammad Ali H. Eljinini and W. M. Hadi, "Nave bayesian based on chi square to categorize arabic data," *Communications of the IBIMA*, vol. 10, no. 20, pp. 158–163, 2009.

[5] G. Forman, I. Guyon, and A. Elisseeff, "An extensive empirical study of feature selection metrics for text classification," *Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, 2003.

[6] M. Porter, "The porter stemming algorithm." `http://tartarus.org/martin/PorterStemmer/`, 2006.

[7] A. Dasgupta, P. Drineas, B. Harb, V. Josifovski, and M. W. Mahoney, "Feature selection methods for text classification," *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining KDD 07*, vol. 21, no. 2, p. 230, 2007.

[8] D. Fragoudis, D. Meretakis, and S. Likothanassis_aff1n3, "Best terms: an efficient feature-selection algorithm for text categorization," *Knowl. Inf. Syst.*, vol. 8, no. 1, pp. 16–33, 2005.

[9] M. Rogati and Y. Yang, "High-performing feature selection for text classification," pp. 659–661, 2002.

[10] A. Kolcz, V. Prabakarmurthi, J. Kalita, and P. Inc, "Summarization as feature selection for text categorization," 2001.

[11] R. Mukras, N. Wiratunga, R. Lothian, S. Chakraborti, and D. Harper, "Information gain feature selection for ordinal text classification using probability re-distribution."

[12] G. Uchyigit and M. Ma, *Personalization techniques and recommender systems*, p. 310. Series in machine perception and artificial intelligence, World Scientific, 2008.

[13] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1289–1305, March 2003.

Fusion Techniques for Satellite Images

Yoonsuk Choi and Shahram Latifi
Dept. of Electrical and Computer Engineering, University of Nevada, Las Vegas
4505 Maryland Parkway, Las Vegas, NV 89154-4026

Abstract
Image fusion involves merging two or more images in a way that the most desirable characteristics of each are retained. When a panchromatic image is fused with a multispectral image, the desired result is an image with higher spatial resolution and more spectral information. Standard image fusion methods are often successful in injecting spatial detail into the multispectral imagery; however, the color information is distorted during the process. Over the past decade, a significant amount of research has been conducted concerning the application of wavelet transforms in image fusion. While wavelet-based schemes are known to perform better than standard schemes, particularly in terms of minimizing color distortion, researchers have been striving to produce even better fusion results by developing contourlet-based schemes and hybrid schemes. This paper is mainly focused on satellite image fusion by reviewing the background of transform theory, and analyzing standard schemes, wavelet-based schemes, contourlet-based schemes, and hybrid schemes.

Keywords: Contourlet, Wavelet, Multispectral, Panchromatic

1. Introduction

Image fusion is a very useful study which can fuse more than one input image to produce an enhanced result by integrating all the positive advantages of the input images. The input images can be acquired from different sources, and they have their own unique advantages, disadvantages and characteristics. There are various areas in the field of image fusion; satellite image fusion for multispectral and panchromatic images, is one of them which has gained remarkable attention over the past few years.

Many efforts have been made to fuse both multispectral and panchromatic images. The standard image fusion techniques based on IHS, PCA, and Brovey transforms usually produce poor results, at least in comparison with the ideal output of the fusion. New fusion schemes have been proposed to address particular problems with the standard techniques; therefore wavelet-based image fusion methods have been developed to better enhance the fusion quality.

Wavelet theory was first applied to signal processing in the 1980's, and over the past decade, it has been recognized as having great potential in image processing applications [1]. Wavelet transforms are essentially extensions of the idea of high pass filtering. Frequency information can be extracted by applying Fourier transforms; however it is no longer associated with any spatial information. Wavelet transforms, on the other hand, are based on functions that are localized in both space and frequency; therefore the wavelet transforms can overcome the disadvantages of Fourier transform. During the fusion process, the extracted detail information from one image using wavelet transforms can be injected into another image using different methods; for example, substitution, addition, or selection method based on either frequency or spatial context. Furthermore, the wavelet function used in the transform can be designed to have specific properties that are useful in the particular application of the transform [2].

Although the wavelet-based fusion schemes produce positive results, they also have drawbacks that must be overcome in the process of image fusion. According to K. Amolins et al. [3], in earlier studies [4]-[6], wavelet-based schemes were generally assessed in comparison to standard schemes; more recent studies propose hybrid schemes [7]-[9], which use wavelets to extract the detail information from one image and inject it into another image, or propose improvements in the method of injecting information [10]-[12]. These approaches seem to achieve better results than either the standard image fusion schemes or standard wavelet-based image fusion schemes [3].

The wavelet transform is good at isolating the discontinuities at object edges, but this transform cannot detect the smoothness along the image edges. Moreover, it can only capture limited directional information. A new contourlet transform has been proposed by Do and Vetterli [13]. The contourlet transform is a new multi-scale, multi-direction framework of discrete image which can capture intrinsic geometric structure information of images and achieve better expression than discrete wavelet transform. Therefore, the contourlet transform can effectively overcome the disadvantages of the wavelet transform.

It may be difficult for readers to understand the transform theories and the current trend of the fusion schemes based on different transforms. The reason is because there are many different areas in the field of image fusion and each area already has various image fusion schemes developed using different transforms. Furthermore, since there is neither a naming convention for these fusion schemes nor an accepted method for assessing the performance of these fusion schemes, it can be difficult to gain a general understanding about existing fusion techniques [3]. Therefore, the purpose of this paper is to facilitate the understanding of satellite image fusion methods,

and to provide a review on each scheme. In Section 2, the principle of wavelet transform is explained briefly to help readers understand the following sections, and the principle of contourlet transform is explained in Section 3 for the same purpose. In Section 4, standard schemes, wavelet-based schemes, contourlet-based schemes, and hybrid schemes are discussed. In the following Sections 5 and 6, the conclusion and future research direction are provided respectively.

2. The Principle of Wavelet Transform

The first wavelet was developed by Alfred Haar in 1909. The Haar wavelet belongs to the group of wavelets known as Daubechies wavelets, which are named after Ingrid Daubechies, who proved the existence of wavelet families whose scaling functions have certain useful properties, namely compact support over an interval, at least one nonvanishing moment, and orthogonal translates. Because of its simplicity (see Equation (1) and Figure 1), the Haar wavelet is useful for illustrating the basic concepts of wavelet theory but has limited utility in applications.

$$\phi_{\text{Haar}}(x) = \begin{cases} 1, & 0 \le x < 1 \\ 0, & \text{otherwise} \end{cases} \quad \psi_{\text{Haar}}(x) = \begin{cases} 1, & 0 \le x < 1/2 \\ -1, & 1/2 \le x < 1 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

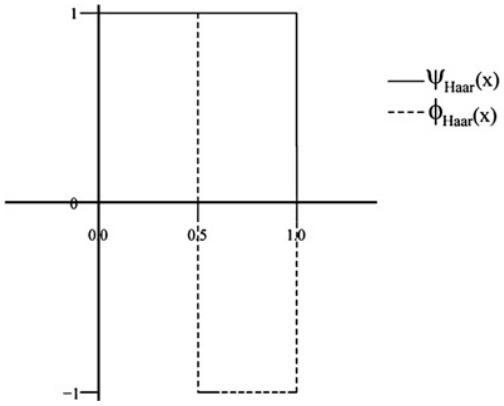


Figure 1: Haar Wavelet

After the first introduction of the wavelet in 1909, more efforts have been made to develop the concept of wavelets. Finally, in 1980's, the relationships between quadrature mirror filters, pyramid algorithms, and orthonormal wavelet bases were discovered, allowing wavelets to be applied in signal processing. Over the past decade, there has been an increasing amount of research on the applications of wavelet transforms in remote sensing image fusion. It has been found that wavelets can be used to extract detail information from one image and inject it into another, since this information is contained in high frequencies and wavelets can be used to select a set of frequencies in both time and space. The resulting fusion image, which can in fact be a combination of any number of images, contains the best characteristics of all the original input images. More comprehensive introductions to wavelets can be found in books, such as [2], [14]-[16].

3. The Principle of Contourlet Transform

The wavelet transform is good at isolating the discontinuities at object edges, but cannot detect the smoothness along the edges. Moreover, it can only capture the limited directional information. The contourlet transform can effectively overcome the disadvantages of wavelet; contourlet transform is a multi-scale and multi-direction framework of discrete image. In this transform, the multi-scale analysis and the multi-direction analysis are separated in a serial way. The Laplacian pyramid (LP) [17] is first used to capture the point discontinuities, then followed by a directional filter bank (DFB) [18] to link point discontinuities into linear structures. The overall result is an image expansion using basic elements like contour segments. The framework of contourlet transform is shown in Figure 2 [19].

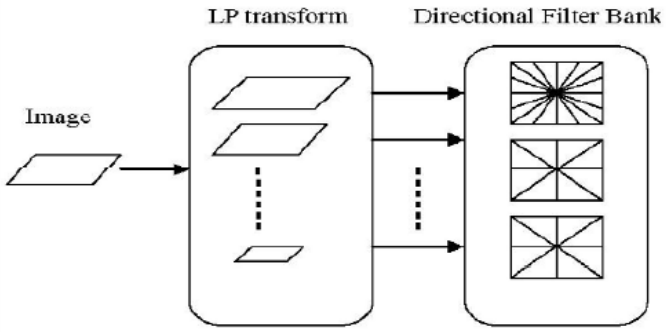


Figure 2: The contourlet transform framework

Figure 3 shows the contourlet filter bank. First, multi scale decomposition is performed by the Laplacian pyramid, and then a directional filter bank is applied to each band pass channel.

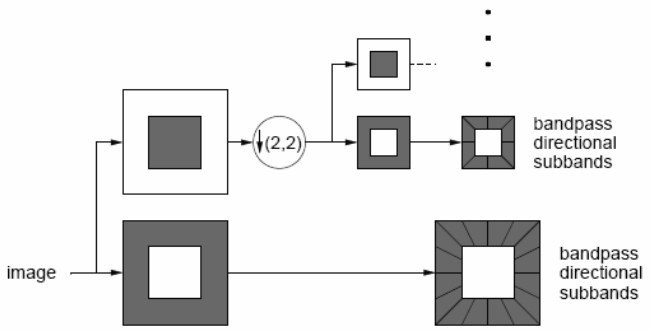


Figure 3: Contourlet filter bank

Contourlet expansion of images consists of basis images oriented at various directions in multiple scales with flexible aspect ratio. In addition to retaining the multi-scale and time-frequency localization properties of wavelets, the contourlet transform offers high degree of directionality. Contourlet transform adopts nonseparable basis functions, which makes it capable of capturing the geometrical smoothness of the contour along any possible direction. Compared with traditional image expansions, contourlet can capture 2-D geometrical structure in natural images much more efficiently [19].

Furthermore, for image enhancement, one needs to improve the visual quality of an image with minimal image distortion. Wavelet bases present some limitations, because they are not well adapted to the detection of highly anisotropic elements such as alignments in an image. Contourlet transform has better performance in representing the image salient features such as edges, lines, curves and contours than wavelet transform because of its anisotropy and directionality.

In order to highlight the difference between the wavelet and contourlet transform, Figure 4 shows a few wavelet and contourlet basis images. It is possible to see that contourlets offer a much richer set of directions and shapes, and thus they are more effective in capturing smooth contours and geometric structures in images.

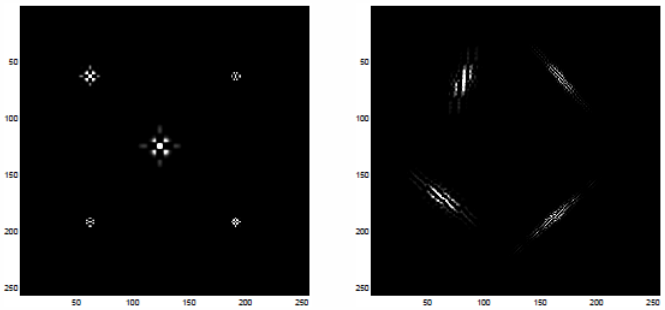


Figure 4: Comparison between actual 2-D wavelets (left) and contourlets (right) [13]

4. Image Fusion Techniques

The objective of image fusion is to produce a single image containing the best aspects of the fused images. Some desirable aspects include high spatial resolution and high spectral resolution (multispectral and panchromatic satellite images), areas in focus (microscopy images), functional and anatomic information (medical images), different spectral information (optical and infrared images), or color information and texture information (multispectral and synthetic aperture radar images). Image fusion can also be used for other purposes, such as providing some protection against illegal copying by embedding watermarks. This paper is mainly focused on the image fusion schemes for satellite images, especially with respect to fusion of multispectral (MS) and panchromatic (PAN) images.

4.1.Standard Fusion Schemes

4.1.1. Standard IHS Fusion

In the IHS color space, intensity (I) is a measure of brightness, with zero representing black, or no brightness, and one representing white, or full brightness. It is also sometimes called luminance (L). Hue (H) is the color, measured as the angle around a color wheel or color hexagon, while saturation (S) is the amount of color, with zero representing grey, or no color, and one representing full color. The main steps of the standard IHS fusion scheme are illustrated in Figure 5 [3].

1. Convert MS from RGB into IHS.
2. Replace I component with PAN.
3. Convert back to RGB.

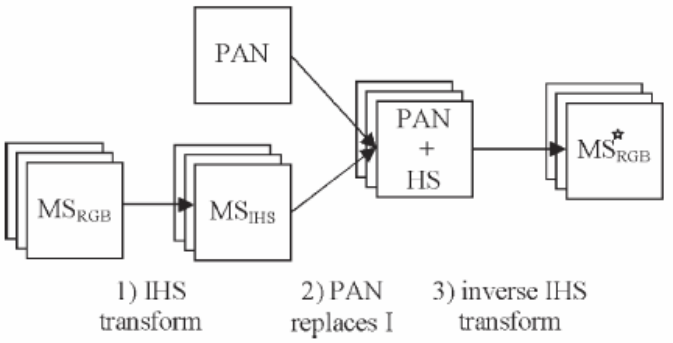


Figure 5: Standard IHS fusion

Various pre-processing steps, such as geometric correction or histogram standardization, can be applied to the images before fusion, in addition to registration and resampling, to improve the fusion results. Furthermore, although there are several different formulas for the IHS transformation, Nunez et al. (1999) [20] demonstrate that the best results can generally be obtained when the intensity is calculated as follows:

$$I = (R + G + B) / 3 \quad (2)$$

One of the drawbacks of the IHS transformation is that it can only be applied to three bands at a time. If there are more than three bands in the MS image, the transformation can be applied to the optimal triplet for each band [3].

4.1.2. Standard PCA Fusion

In most cases, land cover types tend to behave similarly in adjacent bands of the spectrum, causing a significant amount of redundancy in information collected by the sensor. By applying a principle component analysis (PCA), this redundant information can be organized in such a way that each output band is uncorrelated with the others. The first principle component (PC1) contains most of the data variance between all the bands, which, in general, is the spatial information. The PCA fusion scheme is very similar to the IHS fusion scheme, with PC1 being replaced by the PAN image in step 2. Similar to the IHS scheme, pre-processing steps can be applied before the fusion, and histogram matching can be performed between the PAN image and PC1. The PCA fusion scheme has an advantage over the IHS scheme in that it can be applied to all bands in the MS image simultaneously. However, the image must be primarily vegetation to achieve good results because if large areas of the image have very different spectral characteristics, such as water and vegetation, these spectral differences will be contained in PC1 [3].

4.2. Wavelet Based Schemes

Wavelet-based fusion schemes are extensions of the high-pass filter method, which makes use of the idea that spatial detail is contained in high frequencies. In the wavelet-based fusion schemes, detail information is extracted from the PAN image using wavelet transforms and injected into the MS image. Distortion of the spectral information is minimized; however, there may be other negative effects. Some of these effects result from the type of wavelet transform that is used while others result from the method of injecting detail information into the MS image. Various models exist for injection information, with the simplest model being by substitution. Another simple model is by addition, while more complex methods apply mathematical models to the detail images. Regardless of the model, the input images must be at the same resolution in order to be processed for fusion. Depending on the ratio of the original image resolutions, this could necessitate multiple levels of decomposition for the higher resolution image [3].

4.3. Hybrid Wavelet Based Schemes

Various methods have been developed in order to integrate the best aspects of standard methods and wavelet methods. In particular, a great deal of research has focused on incorporating the IHS transform into wavelet methods, since the IHS fusion methods perform well spatially while the wavelet fusion methods perform well spectrally. They can be combined as follows:

1. Convert RGB to IHS.
2. Generate a new PAN image histogram-matched to I component.
3. Apply DWT to new PAN and I images.
4. Apply selected model to obtain set of approximation and detail images.
5. Perform inverse DWT to obtain fused I component.
6. Convert new IHS to RGB.

The model applied in step 4 can be either substitutive wavelet fusion scheme or additive wavelet fusion scheme, with the I component taking the place of the MS image. In other words, the approximation of the I component can directly replace the approximation of the PAN image, the detail images of the I component and the PAN image can be added, a weighting can be applied, or any variation thereof. A similar approach can be used to incorporate the PCA transform with wavelet methods [3].

4.4. Contourlet Based Schemes

4.4.1. Wavelet-based Contourlet Transform (WBCT)

Similar to the contourlet transform, the WBCT consists of two filter band stages. The first stage uses the wavelet transform [21] to provide multi-scale decomposition in contrast to the Laplacian pyramid transform used in contourlet transform. The second stage of the WBCT uses the DFB [21] to provide angular decomposition. The framework of WBCT is shown in Figure 6 [22].

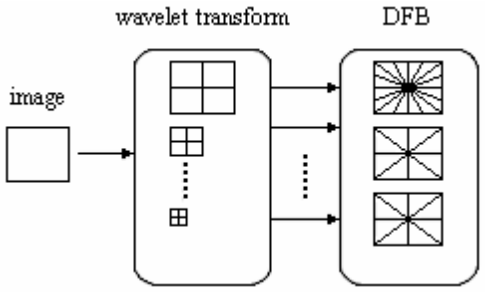


Figure 6: WBCT Framework

As shown in Figure 7, the DFB is designed to capture the high frequency directional components of images. Therefore, the low frequency component is removed before the DFB by the wavelet transform. Then, DFB is applied with the same number of directions to LH, HL and HH bands in a given wavelet level. Starting from the maximum number of directions on the finest level of the wavelet transform, the number of directions at every other dyadic scale is decreased when processing through the coarser levels. Figure 8 illustrates a schematic plot of the WBCT using 3 wavelet levels and 8-4-4 directional levels. The low frequency part is not analyzed by the DFB [22].

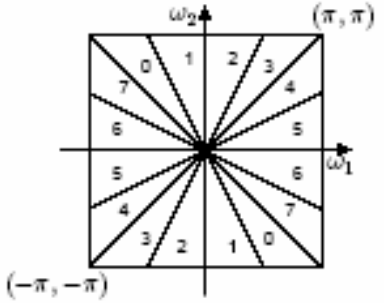


Figure 7: Directional filter bank with 8 frequency bands

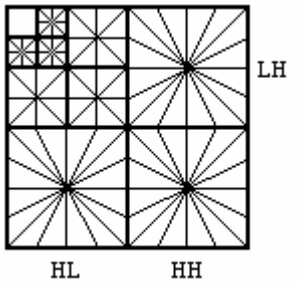


Figure 8: A schematic plot of the WBCT using 3 dyadic wavelet scales and 8-4-4 directions

Another WBCT was implemented by Zhenghua-Shu et al. [23]. The first stage is realized by separable filter banks, while the second stage is implemented using nonseparable filter banks. For the DFB stage, they employ the iterated tree-structured filter banks using fan filters [23]. For each image, the background information belongs to the low frequency part; edge and texture information belongs to the high frequency parts. Therefore, it is possible to use the WBCT to decompose each image into the low frequency parts and high frequency parts; and then different fusion rules can be used in

each part. The WBCT-based image fusion scheme is shown in Figure 9 [23] and the image fusion approach is as follows.

The two images are geometrically registered to each other. First, the WBCT is used in order to decompose the image A and B into multi-scale and multi-direction as shown in Figure 9. Then, the low frequency parts and the high frequency parts are fused together. Finally, the fused image F is reconstructed by performing the inverse of WBCT.

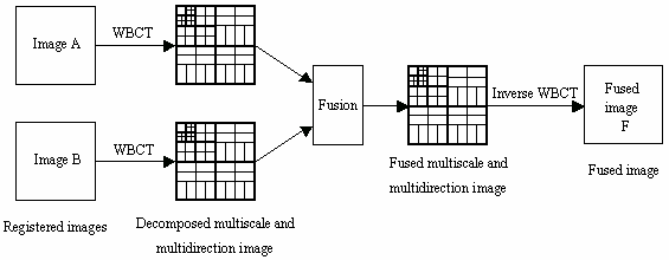
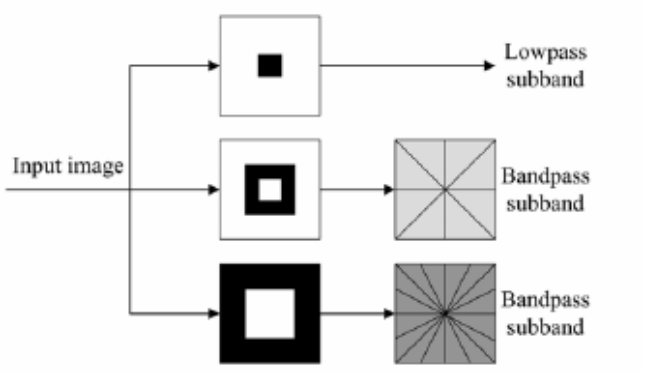


Figure 9: The image fusion framework of WBCT

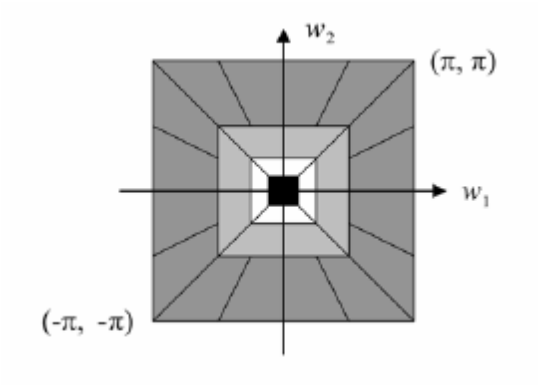
The low frequency part includes most of the background information, and the weighted average operators are used for the fusion rule. In other words, the low frequency coefficients are weighted and averaged. The high frequency parts include most of the image detail information, such as edge and texture information. The correlation between a pixel and its neighboring pixels is often bigger than others, so the fusion rule is about the region where the pixel is at the center. The fusion rules compute the region energy of center pixel and its neighboring pixels.

4.4.2. Nonsubsampled Contourlet Transform (NSCT)

The contourlet transform [13] proposed by Do and Vetterli is a real two-dimensional transform, which is based on nonseparable filter banks and provides an efficient directional multi-resolution image representation. The contourlet transform (CT) expresses image by first applying a multi-scale transform (the Laplacian pyramid transform), followed by a direction filter banks (DFB) to link point discontinuities into linear structures. The contourlets satisfy anisotropy principle and can capture intrinsic geometric structure information of images and achieve better expression than discrete wavelet transform, especially for edges and contours. However, because of the down-sampling and up-sampling, the CT is lack of shift-invariance and results in ringing artifacts. The shift-invariance is desirable in image analysis applications, such as edge detection, contour characterization, image fusion, etc. Cunha et al. proposed nonsubsampled contourlet transform [24] based on nonsubsampled pyramid decomposition and nonsubsampled filter bank (NSFB). In NSCT, the multi-scale analysis and the multi-direction analysis are also separate, but both are shift-invariant. First, the nonsubsampled pyramid (NSP) is used to obtain a multi-scale decomposition by using two-channel nonsubsampled 2-D filter bands. Second, the nonsubsampled directional filter bank is used to split band-pass sub-bands in each scale into different directions. Figure 10 [25] shows two-level decomposition by using a combination of a NSP and NDFB. Because of no down-sample in pyramid decomposition, the low-pass sub-band has no frequency aliasing; even the band width of low-pass filter is larger than $\pi/2$. Hence, the NSCT has better frequency characteristics than CT [25].



(a) NSFB structure that implements the NSCT



(b) Corresponding frequency partition

Figure 10: Two level nonsubsampled contourlet transform decomposition

4.4.3. NSCT + PCA

In the NSCT fusion process, substituting the panchromatic image for high pass bands of multispectral image may lose some important spectral information. A novel hybrid method has been proposed based on a combination of NSCT and PCA. The flowchart of the proposed method is illustrated in Figure 11 [25] and the fusion processing steps are as follows:

Step 1: Perform PCA on the multispectral image A, and get PC1, PC2.

Step 2: Apply histogram matching between the original panchromatic image and PC1 to get approximate mean value and square deviation.

Step 3: Employ NSCT on PC1 and panchromatic image after histogram matching, and get low frequent sub-band and high frequent sub-bands.

Step 4: Fuse the PC1 and panchromatic image. The low frequent data employ low frequent coefficient of PC1. The high

frequent coefficients of panchromatic image are adopted from step 3.

Step 5: Apply NSCT reconstruction with new coefficient to obtain the new PC1.

Step 6: Perform the inverse PCA transform to obtain the fused image.

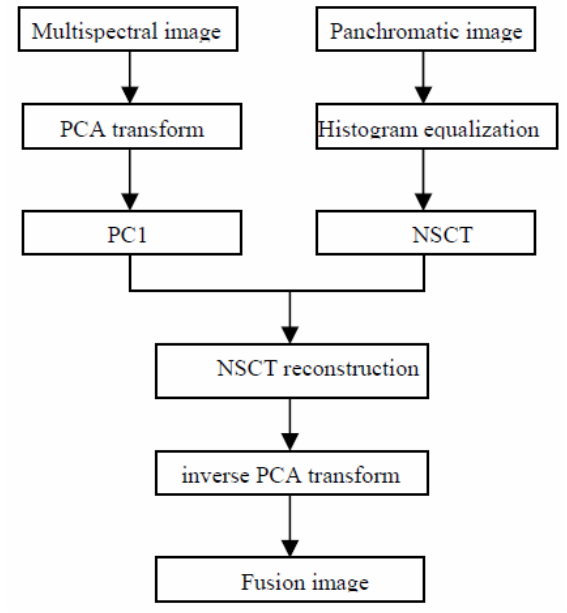


Figure 11: Image fusion flow chart based on NSCT and PCA

4.4.4. IHS + CT

A novel method based on a combination of IHS and CT has been proposed [26]. After applying the contourlet transform to the PAN image and the (I) component, it is possible to obtain their low-frequency and high-frequency sub-images, respectively. Therefore, this newly proposed fusion rule utilizes different fusion rules for different frequencies. In the low-frequency, the improved region-based and PCA weighted fusion rule is applied; and in the high-frequency, the maximum fusion rule is applied.

The proposed fusion scheme is illustrated in Figure 12 [26]. The steps of the proposed new fusion method can be described as follows:

Step 1: Convert the MS image from the RGB color space into the IHS color space, and obtain the H, S, and I components, respectively.

Step 2: Generate a new PAN image by matching the I component with the histogram matching method.

Step 3: Apply contourlet transform to the new PAN and I images.

Step 4: Use the fusion rules mentioned to fuse the low-frequency and high-frequency sub-images, respectively.

Step 5: Perform inverse contourlet transform to obtain the fused I component.

Step 6: Apply the inverse IHS transform to the H, S, and the fused I components, and obtain the final fused image.

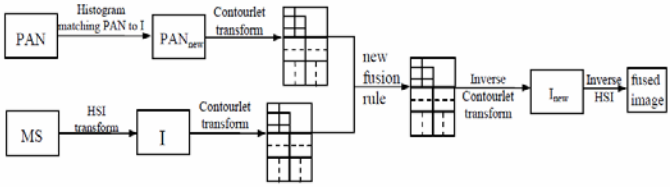


Figure 12: The proposed IHS+CT scheme

5. Conclusion

Image fusion has gained remarkable attention over the past few years, especially in the area of satellite image fusion. A number of different schemes have been proposed to integrate high spatial information of panchromatic imagery with the high spectral information of multispectral imagery. The main purpose of this fusion is to obtain the high spatial resolution while preserving the spectral information.

Various wavelet-based satellite image fusion methods have been developed in order to improve the performance of standard fusion methods such as IHS and PCA. However, the wavelet-based methods have a few drawbacks: greater computational complexity is required; and often parameters must be set up before the fusion scheme can be applied. Another idea for improving the image fusion quality is to combine a standard fusion scheme with a wavelet-based fusion scheme, however this also has a limitation; for example, IHS can only be applied to three bands at a time. Furthermore, the wavelet transform is good at isolating the discontinuities at the object edges, but it cannot detect the smoothness along the edges. As a result, it can capture only limited directional information.

The contourlet transform, a multi-scale and multi-direction framework of discrete image, can effectively overcome the disadvantages of the wavelet. In this transform, the multi-scale analysis and the multi-direction analysis are separated in a serial way. Various contourlet-based fusion schemes are discussed in this paper. NSCT is a potential alternative for satellite image fusion. This scheme performs better in preserving the edge and texture information than that of the other image fusion methods in multi-resolution image fusion field. Also, WBCT scheme is another alternative for satellite image fusion due to its ability to get maximum information entropy, better definition, and contrast. While NSCT and WBCT schemes perform well in fusing the satellite images, researchers have been striving to find a better method by combining two different schemes. In this paper, we have analyzed two hybrid schemes; NSCT+PCA and IHS+CT. Both methods are proven to improve the spatial resolution and hold the largest amount of spectral information.

6. Future Research Direction

Satellite image fusion schemes are mainly focused on obtaining better spatial resolution and more spectral information at the same time. Therefore, people may think that these schemes are only useful in remote sensing field. However, these schemes

can be used in a wide range of applications in various fields, such as battlefield surveillance, target tracking, machine vision, biometrics, and medical diagnosis. Moreover, there are many other new areas that provide room to researchers for further developments. One of them can be hyperspectral imagery processing which is useful in various applications, especially surface surveillance of exposed targets. Unlike conventional single-band or multispectral images, hyperspectral image data are composed of hundreds of contiguous narrow spectral bands. By spatially sharpening a hyperspectral image with a high-resolution panchromatic image, the output image visualization can be enhanced effectively. However, due to the high data dimensionality of hyperspectral data, it is difficult to directly apply hyperspectral images in classification and target detection.

7. Acknowledgment

This work was conducted as a part of an Innovation Working Group supported by the Nevada EPSCoR Programs, and funded by NSF Grant # NSF- EPS-0814372.

8. References

[1] Graps, A., "An introduction to wavelets. IEEE Computational Science and Engineering", 50–61, 1995.

[2] Mallat, S.G., "A Wavelet Tour of Signal Processing", second ed. Academic Press, San Diego, 1999.

[3] Krista Amolins et al., "Wavelet based image fusion techiniques – an introduction, review and comparison", ISPRS Journal of Photogrametry & Remote Sensing, vol. 62, pp. 249-263, 2007.

[4] Garguet-Duport, B., Girel, J., Chassery, J., Pautou, G., "The use of multiresolution analysis and wavelets transform for merging SPOT panchromatic and multispectral image data", Photogrammetric Engineering & Remote Sensing 62 (9), 1057–1066, 1996.

[5] Yocky, D.A., "Multiresolution wavelet decomposition image merger of Landsat Thematic Mapper and SPOT panchromatic data", Photogrammetric Engineering & Remote Sensing 62 (9), 1067–1074, 1996.

[6] Ranchin, T., Wald, L., "Fusion of high spatial and spectral resolution images: the ARSIS concept and its implementation", Photogrammetric Engineering & Remote Sensing 66 (1), 49–61, 2000.

[7] Zhang, Y., Hong, G., "An IHS and wavelet integrated approach to improve pan-sharpening visual quality of natural color IKONOS and QuickBird images", Information Fusion 6 (3), 225–234, 2005.

[8] Gonzalez-Audicana, M., Saleta, J.L., Catalan, R.G., Garcia, R., "Fusion of multispectral and panchromatic images using improved IHS and PCA mergers based on wavelet decomposition", IEEE Transactions onGeoscience and Remote Sensing 42 (6), 1291–1299, 2004.

[9] Gonzalez-Audicana,M., Otazu, X., Fors, O., Seco, A., "Comparison between Mallat's and the 'à trous' discrete wavelet transform based algorithms for the fusion of multispectral and panchromatic images", International Journal of Remote Sensing 26 (3), 595–614, 2005.

[10] Garzelli, A., Nencini, F., "Interband structure modeling for pansharpening of very high-resolution multispectral images", Information Fusion 6 (3), 213–224, 2005.

[11] Otazu, X., Gonzalez-Audicana, M., Fors, O., Nunez, J., "Introduction of sensor spectral response into image fusion methods", Application to wavelet-based methods. IEEE Transactions on Geoscience and Remote Sensing 43 (10), 2376–2385, 2005.

[12] Wu, J., Huang, H., Qiu, Y., Wu, H., Tian, J., Liu, J., "Remote sensing image fusion based on average gradient of wavelet transform", IEEE International Conference on Mechatronics and Automation, 29 July–1 August 2005, pp. 1817–1821, 2005.

[13] M. N. Do and M. Vetterli, "The Contourlet Transform: An Efficient Directional Multiresolution Image Representation", *IEEE Transactions On Image Processing*, vol. 14, no. 12, pp. 2091-2106, Dec. 2005.

[14] Chui, C., "An Introduction to Wavelets", Academic Press, New York 1992.

[15] Prasad, L., Iyengar, S.S., "Wavelet Analysis with Applications to Image Processing", CRC Press, Boca Raton, 1997.

[16] Walnut, D.F., "An Introduction toWavelet Analysis", Applied and Numerical Harmonic Analysis, Birkhauser, Boston, 2002.

[17] Burt, P. J. and E. H. Adelson, "Merging images through pattern decomposition," Proceedings of the SPIE, vol. 575, pp. 173-181, 1985.

[18] Bamberger R H., "A filter bank for the directional decomposition of images: Theory and design", IEEE Trans. Signal Processing, 40 (4): 882 -893, 1992.

[19] Aboubaker M. ALEjaily et al., "Fusion of remote sensing images using contourlet transform", Innovations and Advanced Techniques in Systems, Computing Sciences and Software Engineering, Springer, pp. 213-218, 2008.

[20] Nunez, J., Otazu, X., Fors, O., Prades, A., Pala, V., Arbiol, R., "Multiresolution-based image fusion with additive wavelet decomposition", IEEE Transactions on Geoscience and Remote Sensing 37(3), 1204–1211, 1999.

[21] M. N. Do, "Directional Multiresolution Image Representations, Ph.D. Thesis", *Department of Communication Systems, Swiss Federal Institute of Technology Lausanne*, Nov. 2001.

[22] Lei Tang and Zong-gui Zhao, "The wavelet-based contourlet transform for image fusion", 8th ACIS International Conference on Software Engineering, IEEE, 2007.

[23] Zhenghua-Shu et al., "Remote sensing image fusion based on wavelet based contourlet packet", 2nd Conference on Environmental Science and Information Application Technology, 2010.

[24] Da Cunha A L , Zhou J , Do M N, "The nonsubsampled contourlet transform: theory , design , and applications", IEEE Trans . Image Proc. ,15 (10) : 308923101, 2006.

[25] Ding LI, "Remote sensing image fusion based on nonsubsampled contourlet transform and PCA", International Conference on Computer Tech and Development, 2009.

[26] Mengxin Song et al., "A fusion method for MS and PAN images based on HSI and contourlet transformation", 10th workshop on WIAMIS, pp. 77-80, 2009.

Contourlet Based Multi-Sensor Image Fusion

Yoonsuk Choi and Shahram Latifi
Dept. of Electrical and Computer Engineering, University of Nevada, Las Vegas
4505 Maryland Parkway, Las Vegas, NV 89154-4026

Abstract
Image fusion has brought remarkable attention over the last few years. Images with different specifications, such as resolution, spectral data, and spatial data, can be fused together to produce an image that contains the best features of the input images. Furthermore, it is also possible to fuse images from various sensors like infrared, synthetic aperture radar, panchromatic, multispectral, etc. Wavelet transform has been applied successfully to multi-sensor image fusion, however, due to its drawbacks; new contourlet transform has been used widely since its first introduction in 2005. Therefore, this paper is mainly focused on the multi-sensor fusion methods that are based on the contourlet transform. In this paper, contourlet based multi-sensor image fusion methods are analyzed in four categories: (i) panchromatic and multispectral, (ii) infrared (IR) and visible, (iii) synthetic aperture radar (SAR) and IR, and (iv) SAR and multispectral.

Keywords: Infrared, Multispectral, Panchromatic, SAR

1. Introduction

Various image fusion techniques have been proposed to meet the requirements of different applications, such as concealed weapon detection, remote sensing, and medical imaging. Combining two or more images of the same scene usually produces a better application-wise visible image [1]. The fusion of different images can reduce the uncertainty related to a single image. Furthermore, image fusion should include techniques that can implement the geometric alignment of several images acquired by different sensors. Such techniques are called a multi-sensor image fusion [2]. The output fused images are usually efficiently used in many military and security applications, such as target detection, object tracking, weapon detection, night vision, etc.

The Brovey Transform (BT), Intensity Hue Saturation (IHS) transform, and Principal Component Analysis (PCA) [3] provide the basis for many commonly used image fusion techniques. Some of these techniques improve the spatial resolution while distorting the original chromaticity of the input images, which is a major drawback. Recently, great interest has arisen on the new transform techniques that utilize the multi-resolution analysis, such as Wavelet Transform (WT). The multi-resolution decomposition schemes decompose the input image into different scales or levels of frequencies.

Wavelet based image fusion techniques are implemented by replacing the detail components (high frequency coefficients) from a colored input image with the details components from another gray-scale input image. However, the Wavelet based fusion techniques are not optimal in capturing two-dimensional singularities from the input images. The two-dimensional wavelets, which are obtained by a tensor-product of one-dimensional wavelets, are good in detecting the discontinuities at edge points. However, the 2-D Wavelets exhibit limited capabilities in detecting the smoothness along the contours [4]. Moreover, the singularity in some objects is due to the discontinuity points located at the edges. These points are located along smooth curves rendering smooth boundaries of objects. Do and Vetterli introduced the new two-dimensional Contourlet transform [5]. This transform is more suitable for constructing a multi-resolution and multi-directional expansions using non-separable Pyramid Directional Filter Banks (PDFB) with small redundancy factor [1].

In this paper, we mainly focus on the contourlet transform due to the reason explained above, in order to analyze multi-sensor image fusion methods. The methods are classified into four categories: (i) panchromatic and multispectral, (ii) infrared (IR) and visible, (iii) synthetic aperture radar (SAR) and IR, and (iv) SAR and multispectral. The principle of contourlet transform is briefly explained in Section 2 and the challenges in multi-sensor image fusion are discussed in Section 3. The methods that are classified into four categories are analyzed in Section 4. Lastly, conclusion and future research direction are provided in Section 5 and 6 respectively.

2. The Principle of Contourlet Transform

The wavelet transform is good at isolating the discontinuities at object edges, but cannot detect the smoothness along the edges. Moreover, it can capture limited directional information. The contourlet transform can effectively overcome the disadvantages of wavelet; contourlet transform is a multi-scale and multi-direction framework of discrete image. In this transform, the multi-scale analysis and the multi-direction analysis are separated in a serial way. The Laplacian pyramid (LP) [6] is first used to capture the point discontinuities, then followed by a directional filter bank (DFB) [7] to link point discontinuities into linear structures. The overall result is an image expansion using basic elements like contour segments. The framework of contourlet transform is shown in Figure 1.

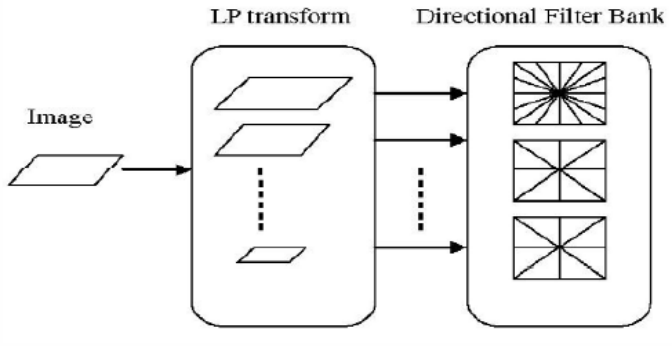


Figure 1: The contourlet transform framework

Figure 2 shows the contourlet filter bank. First, multi scale decomposition by the Laplacian pyramid, and then a directional filter bank is applied to each band pass channel.

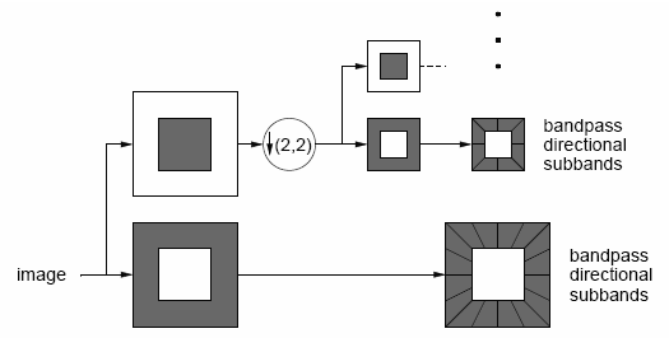


Figure 2: Contourlet filter bank

Contourlet expansion of images consists of basis images oriented at various directions in multiple scales with flexible aspect ratio. In addition to retaining the multi-scale and time-frequency localization properties of wavelets, the contourlet transform offer high degree of directionality. Contourlet transform adopts nonseparable basis functions, which makes it capable of capturing the geometrical smoothness of the contour along any possible direction. Compared with traditional image expansions, contourlet can capture 2-D geometrical structure in natural images much more efficiently [8].

Furthermore, for image enhancement, one needs to improve the visual quality of an image with minimal image distortion. Wavelet-based methods present some limitations because they are not well adapted to the detection of highly anisotropic elements such as alignments in an image. Contourlet transform has better performance in representing the image salient features such as edges, lines, curves and contours than wavelet transform because of its anisotropy and directionality. Therefore, it is well-suited for multi-scale edge based image enhancement.

To highlight the difference between the wavelet and contourlet transform, Figure 3 shows a few wavelet and contourlet basis images. It is possible to see that contourlets offer a much richer set of directions and shapes, and thus they are more effective in capturing smooth contours and geometric structures in images.

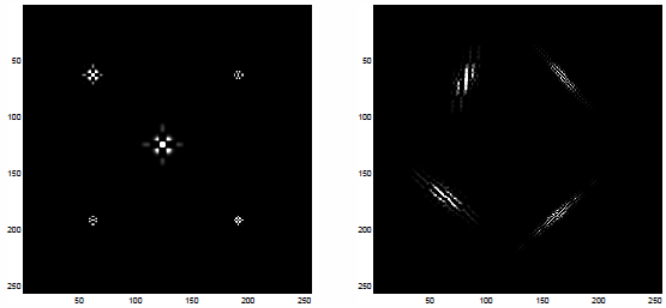


Figure 3: Comparison between actual 2-D wavelets (left) and contourlets (right) [5]

3. The Challenges in Multi-sensor Image Fusion

In the field of image processing, it is necessary for the researchers to establish a direct correspondence between a set of pixels and an object in the image. It is always easier to find and interpret objects from a clear optical picture rather than an obscure picture. However, even if the picture is in high definition, people have hard time finding what they want if the object is covered by clouds, leaves, obstacles or even paints.

In order to perform a better analysis, alternative sensors like microwave and infrared sensors can be used to detect the object attributes that are possibly impossible to be obtained by optical sensors. For example, L-band SAR can find metal objects through the masking of clouds, leaves, tents or paints. Additionally, the infrared imaging devices can reflect the temperature distribution of the object itself and its neighborhood. Usually, the optical image, the radar image, and the infrared image of the same object may be very different in appearance, pattern, and size. In such a case, so-called 'pixel-level image fusion', which aims to improve image resolution, will become lack of scientific ground. Actually, what we need is such a technique that can implement information fusion by making fully use of image information we obtained [2].

In order to find and identify an object reliably, we not only need the object's appearance but also the object's various physical attributes. Figure 4 shows three image samples. Figure 4(a) is a multispectral image with good spectral data. Figure 4(b) is an infrared image of the same area which displays the temperature distribution of the area. Figure 4(c) is a piece of SAR image which shows airplanes. As can be seen, it is hard to distinguish any airplanes in Figure 4(c). However, the ultra-intensive reflection pattern appeared in the image shows the existence of the metal objects. Furthermore, the reflection pattern can offer useful information for confirming the structural characteristics of airplanes [2].

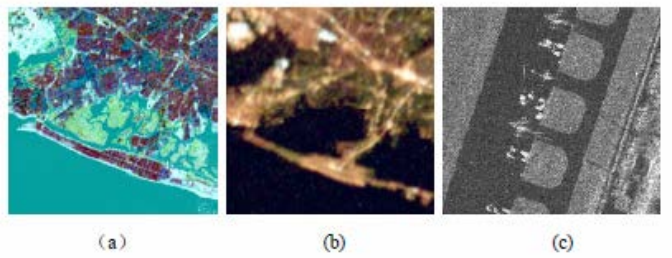


Figure 4: (a) A multispectral image, (b) An infrared image of the same area, (c) A piece of SAR image [2]

In multispectral image processing, images at different wave band of the same vegetation area are usually different. In fact, such difference exhibits the property of the vegetation, i.e. it contains the feature information for distinguishing the property of the vegetation. In order to extract and utilize the information, it is necessary to analyze every image and relate all attributes of each image all together. This is the reason that we should use the fusion method to get all the best positive features of the input images.

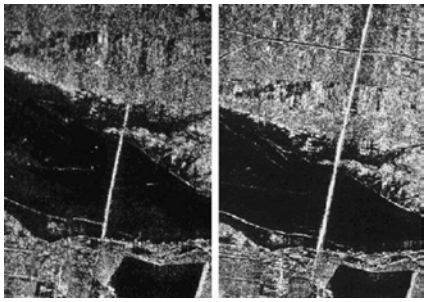


Figure 5: Two polarization SAR images of the same metal bridge (a) VV Polarization and (b) HH Polarization [2]

Polarization SAR image is an important and typical source for the multi-sensor image fusion. The theory and practice of the backscatter of electromagnetic wave show that the reflection will enhance if the orientation of an edge of the object is coincident with the polarization direction of the electromagnetic wave. Especially, reflection will become much intensive if the length of the reflection edge is closer to the electromagnetic wavelength. If SAR can work in HH, VV, HV or VH polarization modes, the acquired images in each polarization mode will be different in general. Figure 5 shows two SAR images of the same metal bridge in different polarization modes. It is possible to see that the length of the bridge in two images looks very different. This difference just reflects the metal attribute of the bridge. Imaging property caused by polarization has evident manifestations not only for metal objects but also for vegetation, soil and so on, although the manifestations may be different. It means that SAR images obtained by various polarizations vary based on the earth surface, i.e. these images contain the feature information of earth surface. This observation has been widely used in agricultural monitoring, land and resource surveying, disaster evaluation and statistics [9]-[11].

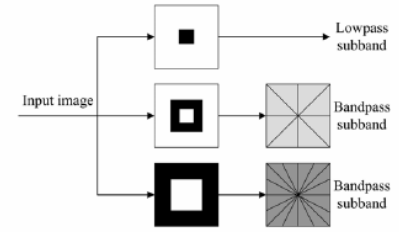
Images acquired from various sensors that are working under different working modes contain a certain particular attributes of the object. These attributes describe the object characteristics from different point of view, such as visual appearance, material, structure size, orientation and temperature; behavior of reflection, radiation and scattering; foreground and background associated with the objects, etc. The purpose of multi-sensor image fusion is to use all the information of each input image to produce a fusion result that can serve as a tool for a certain task.

4. Image Fusion Techniques

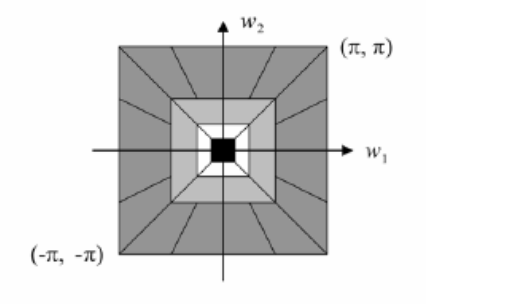
4.1. Panchromatic and Multispectral

4.1.1. Nonsubsampled Contourlet Transform (NSCT)

The contourlet transform [5], [12] proposed by Do and Vetterli is a real two-dimensional transform, which is based on nonseparable filter banks and provides an efficient directional multi-resolution image representation. The contourlet transform (CT) expresses image by first applying a multi-scale transform (the Laplacian pyramid transform), followed by a direction filter banks (DFB) to link point discontinuities into linear structures. The contourlets satisfy anisotropy principle and can capture intrinsic geometric structure information of images and achieve better expression than discrete wavelet transform, especially for edges and contours. However, because of the downsampling and upsampling, the CT is lack of shift-invariance and results in ringing artifacts. The shift-invariance is desirable in image analysis applications, such as edge detection, contour characterization, image fusion, etc. So Cunha et al. proposed nonsubsampled contourlet transform [13] based on nonsubsampled pyramid decomposition and nonsubsampled filter bank (NSFB). In NSCT, the multi-scale analysis and the multi-direction analysis are also separate, but both are shift-invariant. First, the nonsubsampled pyramid (NSP) is used to obtain a multi-scale decomposition by using two-channel nonsubsampled 2-D filter bands. Second, the nonsubsampled directional filter bank is used to split band pass sub-bands in each scale into different directions. Figure 6 shows a two-level decomposition by using a combination of a NSP and NDFB. Because of no downsample in pyramid decomposition, the lowpass subband has no frequency aliasing, even the band width of lowpass filter is larger than $\pi/2$, hence, the NSCT have better frequency characteristics than CT.



(a) NSFB structure that implements the NSCT [14]



(b) Corresponding frequency partition
Figure 6: Two level nonsubsampled contourlet transform decomposition [14]

4.1.2. Wavelet Based Contourlet Transform (WBCT)

Similar to the contourlet transform, the WBCT consists of two filter band stages. The first stage uses the wavelet transform [15] to provide multi-scale decomposition in contrast to the Laplacian pyramid transform used in contourlet transform. The second stage of the WBCT uses the DFB [15] to provide angular decomposition. The framework of WBCT is shown in Figure 7.

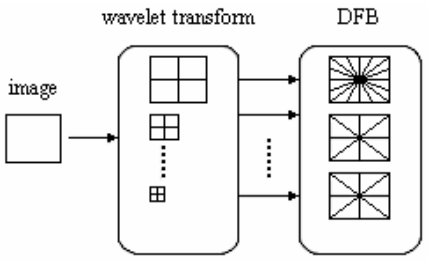


Figure 7: WBCT Framework [16]

As shown in Figure 8, the DFB is designed to capture the high frequency directional components of images. Therefore, the low frequency component is removed before the DFB by the wavelet transform. Then, DFB with the same number of directions is applied to LH, HL and HH bands in a given wavelet level. Starting from the maximum number of directions on the finest level of the wavelet transform, the number of directions is decreased at every other dyadic scale when proceeding through the coarser levels. Figure 9 illustrates a schematic plot of the WBCT using 3 wavelet levels and 8-4-4 directional levels. The low frequency part is not analyzed by the DFB [16].

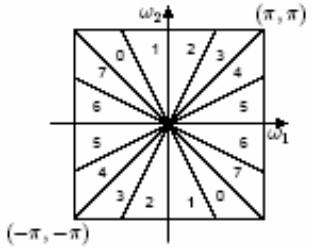


Figure 8: Directional filter bank with 8 frequency bands [16]

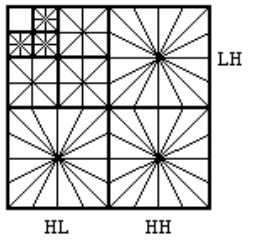


Figure 9: A schematic plot of the WBCT using 3 dyadic wavelet scales and 8-4-4 directions [16]

Another WBCT scheme was implemented by Zhenghua-Shu et al. The first stage is realized by separable filter banks, while the second stage is implemented using nonseparable filter banks. For the DFB stage, they employ the iterated tree-structured filter banks using fan filters [17].

For each image, the background information belongs to the low frequency part, edge and texture information belongs to the high frequency parts. So it is possible to use the WBCT to decompose each image into the low frequency part and high frequency parts, and then different fusion rules can be used in each part.

The WBCT-based image fusion scheme is shown in Figure 10. The image fusion approach is as follows: The two images are geometrically registered to each other. First, use the WBCT to decompose the image A and B into multi-scale and multi-direction as shown in Figure 10. Then, the low frequency parts and the high frequency parts are fused together. Finally, the fused image F is reconstructed by inverse WBCT.

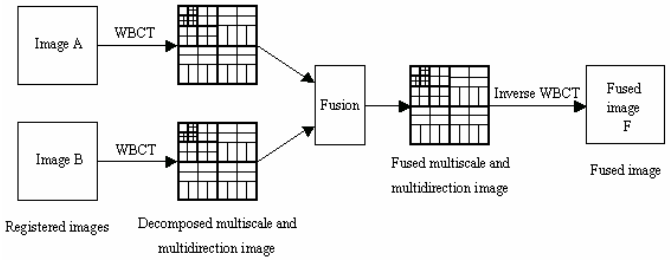


Figure 10: The image fusion framework of WBCT [17]

The low frequency part includes most of the background information, the fusion rules is the weighted average operators. Namely, the low frequency coefficients are weighted averaged. The high frequency parts include most of the image detail information (edge and texture information). The correlation between a pixel and its neighboring pixels is often bigger than others, so the fusion rule is about the region in which the pixel is the center. The fusion rules compute the region energy of center pixel and its neighboring pixels [17].

4.2. Infrared and Visible

Two inputs images are used in this algorithm. The first input image is captured in the visual frequency range; however the second input is an Infrared (IR) image. Generally speaking, many IR sensors use the temperature distribution of the target to produce an IR image [18]. The IR images are usually used for a variety of night-vision applications, such as viewing some

still and moving targets. The infrared radiation, which is emitted from a moving target is absorbed by clothing and then re-emitted. Therefore, the IR image can be used to show the image of the hidden target. Due to the limitation of the IR images in distinguishing all visual objects, the IR images are not sufficient for clearly allocating the target's location. A combination of the IR image with a classical visual image of the same scene can be used to detect the target and precisely identify its location [1].

4.2.1. Intensity-Hue-Saturation (IHS) + Contourlet Transform

The flowchart in Figure 11 summarizes the presented region-based image fusion technique:

1) The two input images (visible colored image and IR gray-scale image) are first co-registered to match their coordinates.
2) The visible colored image is transformed via the IHS transform. The Intensity component (I) is further transformed via the contourlet transform (CT) to yield one of the three inputs for the fusion decision step.
3) Similarly, the IR gray-scale image is also transformed via the CT to yield the second input for the fusion decision step.
4) In the segmentation step, the K-means clustering is used to segment the contourlet coefficients of the IR grayscale image into different regions. The classified output image is the third input for the fusion decision step.
5) The final fusion decision is made based on equation (1), such that, for each region in each decomposition level,

$$Y_F^k = \begin{cases} Y_b^k, & \text{if } Y_b^k \geq Y_a^k \\ Y_a^k, & \text{if } Y_b^k < Y_a^k \end{cases} \quad (1)$$

- The Contourlet coefficients of the fused image are extracted from the IR image if the CT coefficients of the visible image are smaller than the CT coefficients of the IR image.
- The Contourlet coefficients of the fused image are extracted from the visible image if the CT coefficients of the IR image are smaller than the CT coefficients of the visible image.

6) The inverse contourlet transform is applied to produce the new intensity component (I), which contains information from the two input images.
7) The three components: new intensity (I), original H, and original S components are merged via the inverse-HIS transform to produce the output image.

The final output is a colored image combining enough information from both the visible and IR images to identify and precisely locate the hidden target. The final fused image is better for both human and machine interpretations.

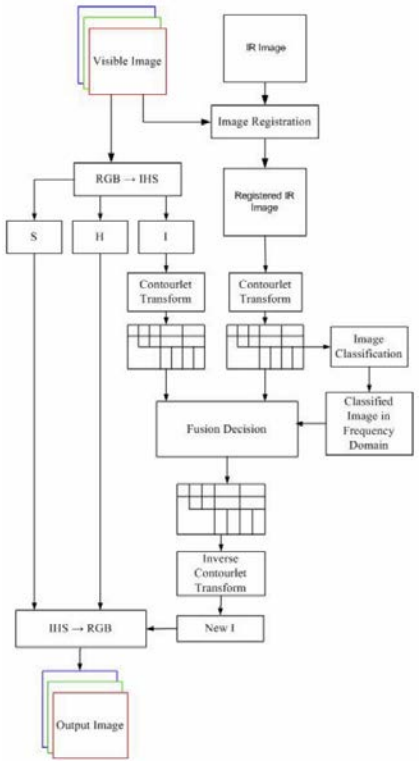


Figure 11: Image fusion process using IHS+CT [1]

4.2.2. Nonsubsampled Contourlet Transform (NSCT)

Nonsubsampled Contourlet Transform (NSCT) is similar with Contourlet Transform, which also do multi-scale and directional decomposition respectively. Firstly, nonsubsampled pyramid filter bank (NSPFB) is used to do multi-scale decomposition with input images. Then, nonsubsampled directional filter bank (NSDFB) is used to do directional decomposition with band pass images of each scale prepared by the first step. Finally, different scale and directional images are obtained [13].

The subsampled procedure after analyzing filtering and upsampled procedure before integrated filtering are removed in laplacian pyramid and directional decompositions of NSCT. These procedures are changed to do upsampled operation to the corresponding filter and the signal is filtered after that. Since it has no upsampled and subsampled procedure, the size of all sub-bands is the same with original images. Because of these reason, NSCT has a character of translation invariance [19].

The proposed algorithm is follows:

1) Suppose the original images are A and B, do NSCT transform respectively to them. A series of directional sub-bands and a low frequency one are obtained.
2) Perform fusion rules to low-frequency sub-band and high-frequency directional sub-bands. The rules could be different to each band to reconstruct image.
3) Reconstruct image using fused coefficients and get the fusion result F.

The low-frequency part contains most of image's energy after multi-scale decomposition, which reflects approximate character of images. Because of gray difference between

infrared and visible images, the targets in these two images often have gray mutual problem. Therefore, infrared and visible image information will be lost by using simple weighted average rule when calculating low-frequency coefficients. In addition, according to the characteristic of human visual system, human eyes are not sensitive to the gray value of one pixel. Furthermore, the degree of identifying target in image is represented by whole pixels of a region. This method uses the rule based on region energy in computing low-frequency coefficients [20], [21] and uses absolute maximum value in calculating high frequency coefficients [19].

4.3. Synthetic Aperture Radar and Infrared

Synthetic Aperture Radar (SAR) is one important branch of radar imaging technology, and it is an active remote sensor system which possesses the ability of all time, all weather, long distance and high resolution. The SAR has been widely used in the field of remote sense, military detection and earth observing. Because the SAR image is radar coherent imaging, it does not have the ability of smoothness along the contours, therefore the SAR image reveals isolating the discontinuities at edge points. The infrared image can reflect approximately the temperature grads and radiation grads of observation object, and it can provide comparatively integrated information of edge and texture.

When SAR image and infrared image are fused, the information of edge and texture obtained from infrared image can be added into SAR image. The edge and texture of fused image will be more integrated, and the frequency characteristics of the SAR image are maintained.

An image fusion algorithm based on the NSCT to fuse SAR image and infrared image is introduced [22]. The fused image combines the good aspects of both SAR and IR images, and as a result, produces an enhanced visual quality.

The fusion algorithm based on NSCT is as follows:

1) Perform a *L*-level NSCT (in this article, L=3) on source images *A* and *B*, and attain the low frequency sub-bands coefficients, and bandpass directional sub-band coefficients.

2) Employ different fusion rules for low frequency sub-band coefficients and bandpass directional sub-bands coefficients to attain the NSCT coefficients of the fused image *AB* .

3) Apply the inverse NSCT on the attained the NSCT coefficients and thus obtain the fused image *AB* .

The simulated results show that the NSCT is suitable for fusing SAR and IR images. Furthermore, the results demonstrate that higher performance and better visual quality can be achieved compared to the discrete wavelet transform (DWT) based fusion method [22].

4.4. Synthetic Aperture Radar and Multispectral

The scene edges in SAR images are incomplete; hence, the texture of the image's edge is discrete. Therefore, there are strong speckle noises. These images are different from the images that are accepted by the human visual system, and this difference has brought great difficulties to the interpretation of images. As a result, the fusion of SAR and multispectral images can be one of the solutions to improve the visualization which in turn provides both visual and spectral data [23].

A new method of fusing SAR and multispectral images based on contourlet-IHS (CT-IHS) has been proposed [23]. Although contourlet transform (CT) could be employed lonely to fuse images, after images are decomposed with CT, high-frequency component and approximation are integrated. Then perform an inverse CT to finish fusion. The detailed steps of this integrated fusion method are as follows:

1) Reduce the speckle of SAR image using Lee filter.

2) Transform the multispectral image into IHS space.

3) Apply histogram of I component to match the histogram of de-speckled SAR image.

4) Decompose both the new I component and the SAR image using CT; three-level decomposition is applied.

5) Replace the approximation image of I component using that of the SAR image. In order to improve definition of fusion result, the maximum amplitude value of the SAR image and I component are selected as new detail components.

6) Perform an inverse CT to obtain a new intensity image (new I' component).

7) Transform the new I' component together with the hue and saturation components back into RGB space.

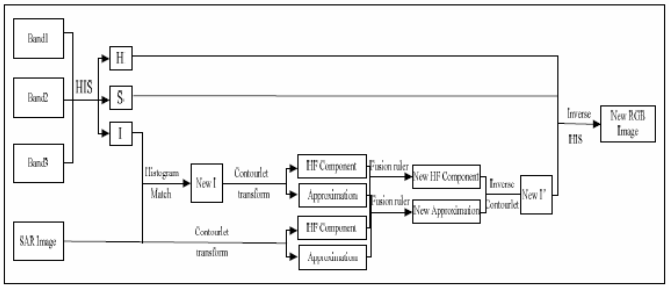


Figure 12: Fusion process of the proposed CT-HIS method [23]

The proposed method makes use of the advantages of contourlet transform; CT is able to capture spatial details efficiently, and integrate IHS transform to fuse SAR images and multispectral images. The evaluation shows that the performance of the fusion result based on CT-IHS transform is better than the ones based on stand-alone IHS and DWT-IHS methods in terms of integrating spectral characteristics of multispectral images and spatial details of SAR images [23].

5. Conclusion

Novel multi-sensor image fusion methods are discussed by analyzing four different sensor categories. Our study mainly focuses on the fusion methods that are based on the contourlet transform due to the following observation. Various methods have been proposed for multi-sensor image fusion; however, both standard transforms (IHS, PCA, DWT) and wavelet transform cannot detect the smoothness along the edges. As a result, the spatial details of the image are not described

efficiently. On the other hand, contourlet transform is a new multi-scale and multi-direction image analysis method which can effectively overcome the disadvantages of the wavelet-based methods. All the methods discussed in this paper perform better than the wavelet-based methods.

6. Future Research Direction

Multi-sensor image fusion has recently gained much attention due to its usefulness in various applications. The study of this specific area of image fusion has only begun, especially after the introduction of the contourlet transform in 2005. From now on, researchers should develop novel methods for fusing different combinations of sensors, such as multispectral and infrared, SAR and hyperspectral, etc. Also, there are many more visual sensors that are not discussed in this paper like medical sensors; for example, magnetic resonance imaging (MRI) and computed tomography (CT). Furthermore, hyperspectral imaging is another future challenge in the area of multi-sensor image fusion. The opportunity for hyperspectral data fusion occurs in surface surveillance of exposed targets. Unlike conventional single-band or multispectral sensors, hyperspectral sensor collects image data in hundreds of contiguous narrow spectral bands. By spatially sharpening a hyperspectral image with a panchromatic high-resolution image, we can enhance the image visualization effectively.

7. Acknowledgment

This work was conducted as a part of an Innovation Working Group supported by the Nevada EPSCoR Programs, and funded by NSF Grant # NSF- EPS-0814372.

8. References

[1] S. Ibrahim and M. Wirth, "Visible and IR Data Fusion Technique Using the Contourlet Transform", International conference on computational science and engineering, CSE 09, IEEE, vol. 2, pp. 42-47, 2009.

[2] Mouyan Zou and Yan Liu, "Multi-Sensor Image Fusion: Difficulties and Key Techniques", 2nd International congress on image and signal processing, IEEE, pp. 1-5, 2009.

[3] G. Pajares and M. de la Cruz, "A wavelet-based image fusion tutorial," Pattern Recognition, vol. 37, no. 9, pp. 1855–1872, 2004.

[4] G. V. Welland, Beyond Wavelets. Academic Press, 2003.

[5] M. N. Do and M. Vetterli, "The contourlet transform: An efficient directional multiresolution image representation," IEEE Transactions on Image Processing, vol. 14, no. 12, pp. 2091–2106, 2005.

[6] Burt P J., "Merging images through pattern decomposition", Proceedings of SPIE, 575: 173-18, 1985.

[7] Bamberger R H., "A filter bank for the directional decomposition of images: Theory and design", IEEE Trans. Signal Processing, 40 (4): 882 -893, 1992.

[8] Aboubaker M. ALEjaily et al., "Fusion of remote sensing images using contourlet transform", Innovations and Advanced Techniques in Systems, Computing Sciences and Software Engineering, Springer, pp. 213-218, 2008.

[9] Rignot, E. Chellappa, R., Segmentation of Polarimetric Synthetic Aperture radar Data, IEEE Trans. on Image Processing, 1(3): 281 – 300, 1992.

[10] Gustavo Camps-Valls, Luis Gómez-Chova, Jordi Muñoz-Marí, et al., Kernel-Based Framework for Multitemporal and Multisource Remote Sensing Data Classification and Change Detection, IEEE Transactions on Geoscience and Remote Sensing, 46(6): 1822-1835, 2008.

[11] Henrik Aanæs, Johannes R. Sveinsson, Allan Aasbjerg Nielsen, et al., Model-Based Satellite Image Fusion, IEEE Transactions on Geoscience and Remote Sensing, 46(5): 1336-1346, 2008.

[12] Do M N, Vetterli M., "Contourlets: A directional multiresolution image representation", Proc of IEEE International Conference on Image Processing, Rochester, NY, pp. 357-360, 2002.

[13] Da Cunha A L , Zhou J , Do M N, "The nonsubsampled contourlet transform: theory , design , and applications", IEEE Trans . Image Proc. ,15 (10) : 308923101, 2006.

[14] Ding LI, "Remote sensing image fusion based on nonsubsampled contourlet transform and PCA", International Conference on Computer Tech and Development, 2009.

[15] M. N. Do, "Directional Multiresolution Image Representations, Ph.D. Thesis", *Department of Communication Systems, Swiss Federal Institute of Technology Lausanne*, Nov. 2001.

[16] Lei Tang and Zong-gui Zhao, "The wavelet-based contourlet transform for image fusion", 8th ACIS International Conference on Software Engineering, IEEE, 2007.

[17] Zhenghua-Shu et al., "Remote sensing image fusion based on wavelet based contourlet packet", 2nd Conference on Environmental Science and Information Application Technology, 2010.

[18] H.-M. Chen, S. Lee, R. M. Rao, M.-A. Slamani, and P. K. Varshney, "Imaging for concealed weapon detection," IEEE Signal Processing Magazine, pp. 52–61, 2005.

[19] Huang Qingqing et al., "Improved Fusion Method for Infrared and Visible Remote Sensing Imagery Using NSCT", 6th IEEE conference on industrial electronics and applications, pp. 1012-1015, 2011.

[20] Shapiro J M. Embedded Image Coding Using Zerotrees of Wavelet Coefficients[J]. IEEE Trans Acoust Speech Signal Processing, ASSP-36(9) : 1445-1453, 1988.

[21] Yajun Song, Guoqiang NI, et al. Regional Energy Weighting Image Fusion Algorithm by Wavelet Based Contourlet Transform[J]. Transactions of Beijing Institute of Technology, 28(2) :168-172, 2008.

[22] Ying Zhang et al., "SAR and Infrared Image Fusion Using Nonsubsampled Contourlet Transform", JCAI 09, IEEE, pp. 398-401, 2009.

[23] Dengshan Huang et al., "SAR and Multi-spectral images fusion based on Contourlet and HIS Transform", WICOM, IEEE, pp. 1-4, 2010.

Investigating the Relationship between Neonatal mortality rate and Mother's characteristics

M. Abdollahian[*], S. Ahmad[#], S. Huda[+], S. Nuryani, D.Anggraini[y]
[*]*School of Mathematical and Geospatial Sciences, RMIT University, Melbourne, Australia*
#Department of Management, College of Business Administration Al Yammamah University, Riyadh Saudi Arabia
[+]*School of Information Technology and Mathematical Sciences, University of Ballarat, Australia*
[x]*Ulin Hospital (RSUD Ulin) Banjarmasin Indonesia*
[y]*Department of Mathematics Lambung Mangkurat University, Banjararu, Indonesia*

Abstract

*Neonatal mortality rate **(NMR)** is an increasingly important public health issues in many developing countries. Neonatal death now accounts for about two-thirds of the eight million infant deaths that occur globally each year. It is well-documented that low birth weight **(LBW)** is the most significant factor influencing NMR. This paper deploys regression analysis to explore the relationship between weight of low birth weight babies and various characteristics of mother. The results indicate that there is a significant relationship between weight of low birth weight babies and mother's weight, age, gestation age and hemoglobin level.*

Keywords: Normality test, multivariate normal distribution, simulation and multi-regression

1. Introduction

Newborn size is an important indicator of infant survival and childhood mortality. Simple and accurate method of estimating newborn weight that can be easily applied to all pregnancies is an important means of reducing mortality rate. Several investigators have shown that low birth weight is associated with high prenatal mortality and morbidity [6]. Birth weight may also predict both short- and long term adverse outcomes. For example, higher birth weight among term infants is associated with birth complications [2] as well as a reduced risk of cardiovascular disease and hypertension in later life, but an increased risk of obesity [7-10].

The ultrasound has been used for the examination and evaluation of high-risk pregnancies and for the diagnosis of congenital malformations. During the last two decades ultrasound techniques have been improved and are implemented in most gynecology and obstetric clinics worldwide [11]. The biophysical profile was used to assess fetal well-being and to confirm fetal gestational age by measuring the biparietal diameter and crown rump length [12]. Other investigators have predicted intrauterine fetal weight using ultrasonographic measurement of the fetal abdominal circumference [3]. More recent reports have emphasized the usefulness of this measurement in monitoring normal fetal growth and in detecting intrauterine growth retardation [4]. All these studies were conducted in Western countries where perinatal medical care and ultrasonographic measurements are advanced.

This study was undertaken among pregnant women attending the Banjarmasin Clinic in Indonesia.during 2010-2011. A total of 198 pregnant women between the ages of 16 and 42, who attended the clinic for antenatal care or routine follow up were included in the study. For each patient; age, baby weight, patient weight at the time of delivery (Kg), gestation age at the time of delivery, hemoglobin level before (in the third semester of pregnancy) and after (towards the end after consumption of Vitamin C and Sulfas Ferroses*)* were measured. Out of the 198 deliveries 10 had a low birth weight. Low birth weight is defined as a birth weight of less than 2500 gram and is a well-documented risk factor for neonatal mortality [1, 5]. Information on these characteristics for individual babies was obtained from the records. In this study we have investigated the distribution of all these characteristics for the low weight babies. It was observed that all characteristics follow normal distribution. Multivariate normal distribution based on the observed means and standard deviations is used to obtain 1000 simulated data. Multi-regression analysis is deployed to find the relationship between weight for low weight birth babies and the above characteristics.

This paper is organized in the following manner. Time series and distribution analysis are discussed in section 2. A review

of the multi-regression analysis and multi-normal stimulations are presented in section 3. Discussion based on Simulated and application example with real clinical data is presented in section 4. This followed by conclusion in section 5.

2. Time series and Distribution analysis

2.1 Time series plots are used to evaluate patterns in data over time. It is also used to investigate whether the patterns on different characteristics have uniform or corresponding pattern. For our experimental data, we first decided to compare the time series pattern of the detailed characteristics by including all 198 patients. The plot for the first 30 patients is presented in Figure 1.

The result clearly shows that the trend in baby' weight follows the trend in mother age, digestion age, hemoglobin level and mother weight. However, the variability in baby's weight almost consistently matches the trends present in the mother's age and hemoglobin level; i.e., when the hemoglobin level is down the baby weight is also down.

2.2 Normality test

To carry out statistical analysis one often requires normal data. The normal probability plot is a graphical technique for assessing whether or not a data set is approximately normally distributed. The authors have used statistical used statistical package Minitab to fit the commonly used normal distribution function to individual characteristic of the low birth weight data. For all the characteristics in the data set the p-value of the fit was less than 0.01 indicating that all the six characteristics follow normal distribution. The graph is presented in Figure 2.

We have used the multi-normal distribution based on the means and standard deviations of these fitted individual normal distributions to generate one thousand samples of size one with six characteristics in each sample. The Normal probability plots for the simulated data together with the corresponding p-value for the individual characteristics are given in Figure 3.

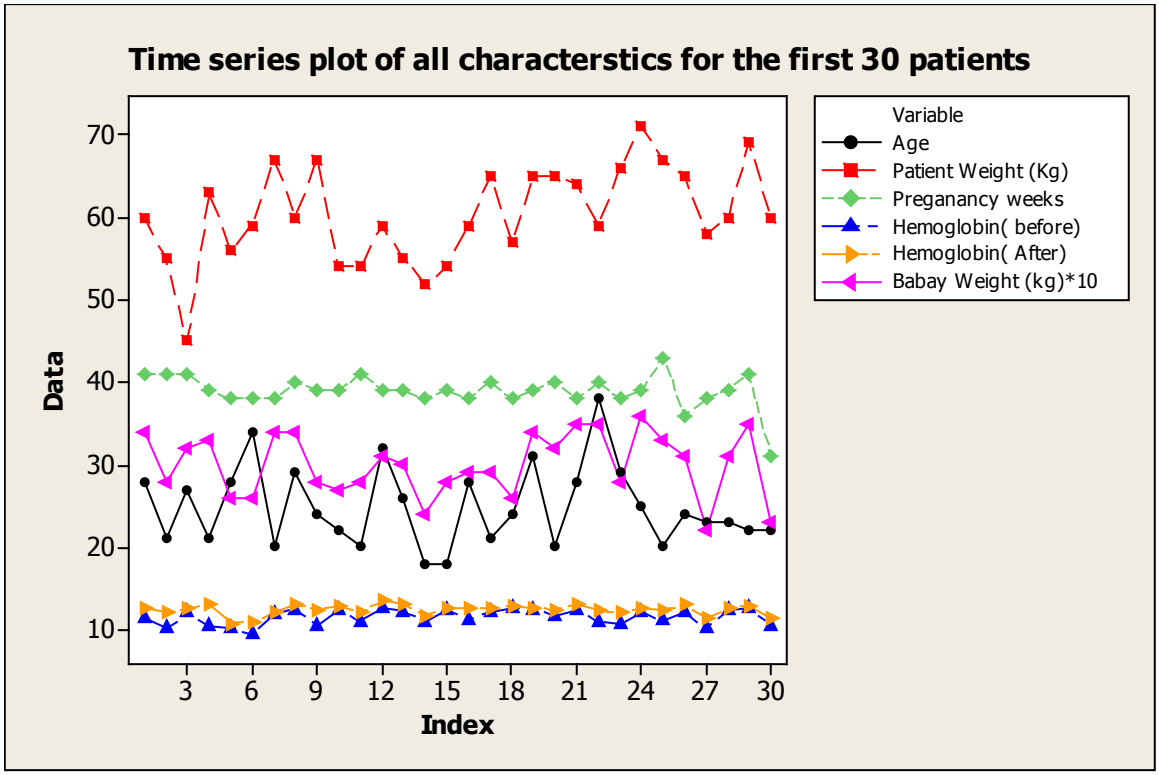


Figure 1: Time series plot of all the six characteristics

.

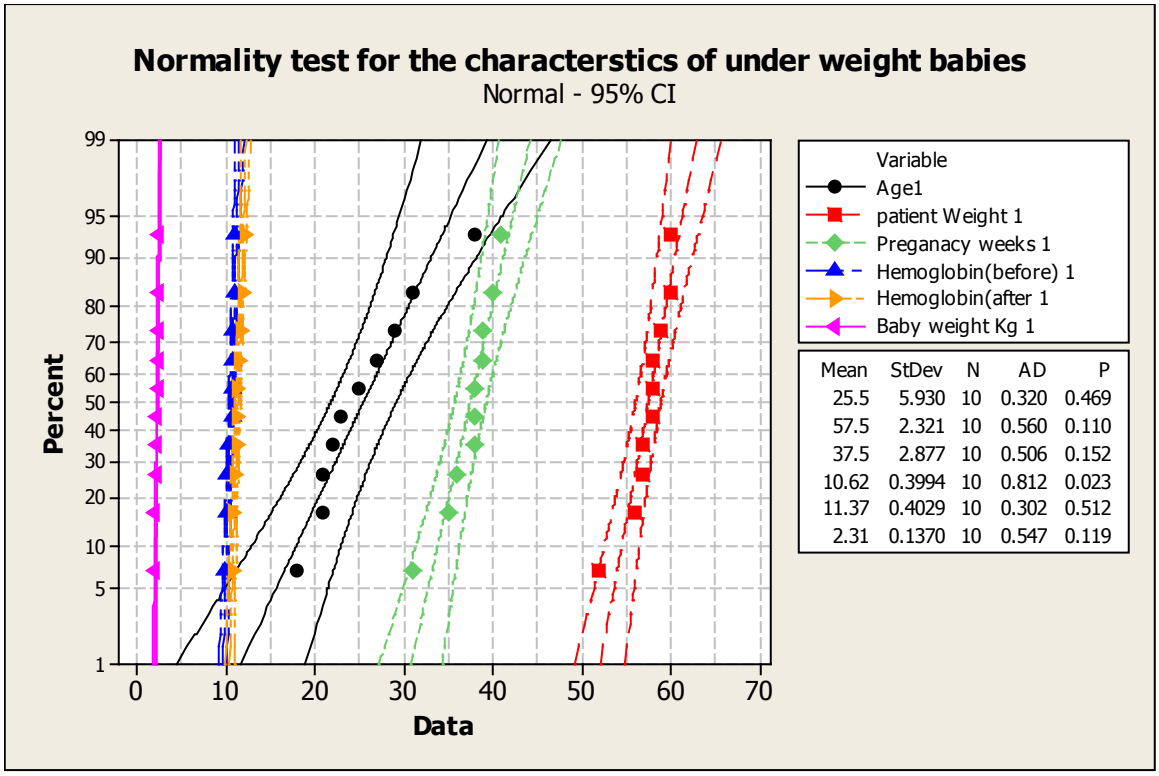


Figure2: Normal probability plot of the actual data, where the p-value for each characteristic is listed.

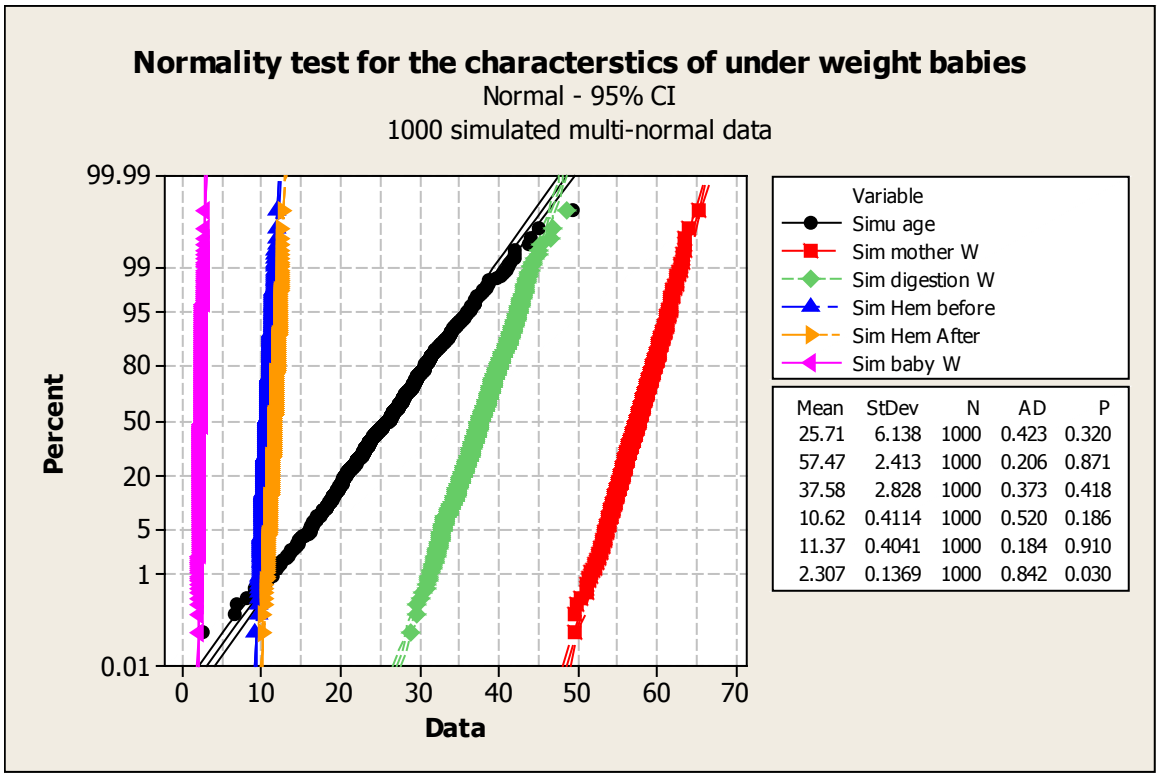


Figure 3: Normal probability plot of the 1000 simulated data with the corresponding p-value for individual characteristic.

Figure 3 shows that the p-value corresponding to all characteristics is greater than 0.01. Therefore we can conclude that the 1000 simulated data follow normal distribution.

3. Multi-Regression Model and Multi-normal Simulation

Multiple-regression is an appropriate approach used to accurately model the relationship between a scalar variable *y* and one or more explanatory variables denoted by *X*. It is assumed in multiple regressions that the residuals (predicted minus observed values) are distributed normally. Even though most tests (specifically the *F*-test) are quite robust with regard to violations of this assumption, it is always a good idea, before drawing final conclusions, to review the distributions of the major variables of interest.

Given a data set $\{y_i, x_{i1}, x_{i2}, \ldots, x_{ip}\}$ of *n* statistical units, a linear regression model assumes that the relationship between the dependent variable y_i and the *p*-vector of regressors x_i is linear. This relationship is modelled through a so-called "disturbance term" ε_i , an unobserved random variable that adds noise to the linear relationship between the dependent variable and regressors. Thus the model takes the form

$$y_i = \beta_1 x_{i1} + \cdots + \beta_1 x_{ip} + \epsilon_i \quad i = 1, \ldots, n. \qquad (3.1)$$

In this study the authors are interested to find the relationship between; the baby weight for low birth weight babies (dependent variable) and mother age, mother weight , gestation age at the time of delivery and hemoglobin level (before and after). The independent effect of a number of variables on the baby weight was calculated by using multiple linear regressions. Using 1000 multi-normal simulated data based on the means and standard deviations of the normal distribution fitted to individual observed characteristics, we obtained the following regression model:

Baby W = 1.46 - 0.00126 age + 0.00299 mother W + 0.00483 digestion W + 0.263 Hem before - 0.200 Hem After (3.2)

The correlation coefficient r = 0.8 for the regression equation given in (3.2).

To assess the efficacy of the proposed model, we have deployed the model to predict the observed weight of the actual data. The predicted value, together with the standard error of the prediction, observed and predicted 95% confidence interval, error of prediction and the actual observed weight are given in table 1

Table 1: Predicted Values for the actual baby weight based on the proposed multi regression model.

New Obs	Fit	SE Fit	95% CI	95% PI	Error of prediction	Actual Value
1	2.35634	0.00717	(2.34228, 2.37040)	(2.19305, 2.51962)	0.04366	2.4
2	2.17726	0.00688	(2.16375, 2.19077)	(2.01402, 2.34050)	0.02274	2.2
3	2.24332	0.00716	(2.22927, 2.25736)	(2.08003, 2.40661)	0.05668	2.3
4	2.23267	0.00513	(2.22261, 2.24273)	(2.06968, 2.39566)	0.06733	2.3
5	2.22985	0.00580	(2.21848, 2.24122)	(2.06677, 2.39293)	-0.12985	2.1
6	2.44998	0.00607	(2.43806, 2.46190)	(2.28686, 2.61310)	-0.04998	2.4
7	2.27899	0.00548	(2.26824, 2.28974)	(2.11595, 2.44202)	0.12101	2.4
8	2.44922	0.00557	(2.43830, 2.46015)	(2.28617, 2.61227)	0.05078	2.5
9	2.45621	0.00509	(2.44622, 2.46620)	(2.29322, 2.61920)	-0.05621	2.4
10	2.19371	0.00636	(2.18123, 2.20618)	(2.03055, 2.35687)	-0.09371	2.1

Table 1 presents the predicted weight in column 2 under "Fit", corresponding 95% confidence interval under "95% CI", Actual observed weight which is listed in the last column under "Actual value" and error of the prediction based on the proposed model under "Error of prediction". It can be seen that the maximum forecasting error for the proposed model corresponds to observation 7 which is 121 grams.

Table 2: Summary statistics for the forecasting error of the proposed model

N	Mean	SE Mean	StDev
10	0 0.0032	0.0255	0.0807

Table 2 shows that the proposed regression model can predict the true value of the new borne weight for LBW babies with the mean error of 0.0032 (3.2 g) and the standard

error of 0.0255. Therefore we can claim that the accuracy of the model is significant and precise.

4. Discussion

The study was conducted among women enrolled in a maternity clinic in Banjarmasin Indonesia. One of the objectives in this research experiment was to investigate the relationship between baby weight (for the low birth weight babies) and mother age, weight, gestation age at the time of delivery, hemoglobin level (before and after consumption of Vitamin C and Sulfas Ferroses) and propose a model to predict weight of LBW babies based on theses characteristics. The independent effect of a number of variables on the baby weight was calculated by using multi-regressions. The slope (the beta-coefficient) that shows the amount of change in the dependent (baby weight) for one unit change in an independent variable, such as, mother weight, and other variables together with the Pearson correlation coefficient were used to identify the effect of independent variables on the baby weight. All the analyses were performed using the MINTAB statistical software package.

The slops and their corresponding p-values in the proposed regression model show that the most effective variables are the gestation age (slope of .005), hemoglobin level before (slope = 0.263)and after (slope = -0.2) and the least effective variables are mother weight (slope = 0.003)and age (slope = -0.00126).

The proposed regression model has a correlation coefficient of 80%. Therefore, one can claim that the relationship is strong enough to predict the newborn weight based on the other four independent variables. The model was then used to predict the observed sample weight data to assess its predicting accuracy. Results presented in tables 1 and 2 shows that the estimated weight using the proposed model is very close to the actual recorded weight for the LBW babies with the mean error prediction of 3.2 grams.

5. Conclusion

Low birth weight is an increasingly prevalent factor in the Maternal Mortality Rate (MMR). Therefore many studies have attempted to identify the sources of variation in the newborn weight. In this study, multi-regression model is used to assess the independent effects of the mother age, weight, gestation age at the time of delivery and hemoglobin level (before and after) on the new born weight for low weight babies.

One thousand Multi-normal simulated data based on the means and standard deviations of the recorded low birth weight newborns were used to estimate the model. The results show that for low birth weight babies there is a statistically significant relationship between the new born weight and the independent variables; mother age, weight, gestation week and hemoglobin level with a correlation coefficient of 80%.

The proposed model was used to estimate the recorded weights together with their corresponding 95% confidence interval. Analysis of the prediction errors shows that the mean prediction error for the recorded data is 3.2 grams. Therefore one can conclude that the proposed multi-regression model is capable of accurately predicating the weight for the low birth weight babies based on the characteristics of the mother. The model is based on one thousand simulated data using the sample measurements and in future would be tested on a larger set of observed data.

Acknowledgement

The authors would like to thank the medical practitioners and staff in the maternity clinic in Banjarmasin Indonesia for their effort in collecting this data and providing them to us for this research.

References

[1] Allen M, Mor J. US birth weight/gestational age-specific neonatal mortality: 1995-1997 rates for whites, Hispanic, and blacks. Pediatrics 2003;111:61-6.

[2] Bennett BB. Shoulder dystocia: An obstetric emergency. Obstet Gynecol Clin North Am 1999;26:445-58.

[3] Campbell S, Wilkin D. Ultrasonic measurement of fetal abdominal circumference in the estimation of fetal weight. Br J Obstet Gynecol 1975;82:689-97.

[4] Chang TC, Robson SC, Boys RJ, Spencer JA. Prediction of the small for gestational age infant: Which ultrasonic measurement is the best?.Obstet Gynecol 1992;80:1030-7.

[5] Mathews TJ, MacDorman MF, Menacker F. Infant mortality statistics from the 1999 period linked birth/death data set. Natl Vital Stat Rep 2002;50:1-28.

[5] Mavalankar DV, Trived CR, Gray RH. Levels and risk factors for prenatal mortality in Ahmadabad, India. Bull World Health Organ 1991;69:35-42.

[7] Oken E, Gillman MW. Fetal origins of obesity. Obstet Res 2003;11:496-506.

[8] Phillips DI. Birth weight and the future development of diabetes: A review of the evidence. Diabetes Care 1998;21:150-5.

[9] Rich-Edwards JW, Stampfer MJ, Manson JE, Rosner B, Hankinson SE, Colditz GA, et al. Birth

weight and risk of cardiovascular disease in a cohort of women followed up since 1976. BMJ 1997;315:396-400.

[10] Rich-Edwards JW, Colditz GA, Stampfer MJ, Willett WC, Gillman MW, Hennekens CH, et al. Birth weight and the risk for type 2 diabetes mellitus in adult women. Ann Intern Med 1999;130:278-84.

[11] Saari-Kemppainen A, Karjalainen O, Ylostalo P, Heinonen OP. Ultrasound screening and perinatal mortality: Controlled trail of systematic one-stage screening in pregnancy: The Helisinki Ultrasound Trail. Lancet 1990;336:387-91.

[12] Skovron ML, Berkowitz GS, Lapinski RH, Kim JM, Chitkara U. Evaluation of early third-trimester ultrasound screening for intrauterine growth retardation. J Ultrasound Med 1991;10:153-9.

Multivariate Exponentially Weighted Moving Average chart for monitoring patient's progress after cardiac surgery

M. Abdollahian and P. Hayati Rezvan
School of Mathematical and Geospatial Sciences, RMIT University, Melbourne, Victoria, Australia

Abstract- *Statistical process control has emerged in the medical literature after wide expansion in the industry. In clinical monitoring, there are always more than one quality characteristics of interest which are usually correlated. In such cases, multivariate control charts would be deployed to monitor the medical process. In this paper, Multivariate Exponentially Weighted Moving Average control chart (MEWMA) is applied to monitor the patient's progress in the Intensive Care Unit, which is characterised by nine quality characteristics. One difficulty encountered with multivariate control charts is the interpretation of out-of-control signals. The univariate control charts are employed to obtain a rough estimate of the sources of multivariate out-of-control signals. Issues of non-normality in the data are addressed,and suitable transformations are offered.*

A comparison is made between the performance of EWMA and MEWMA methods in monitoring of patient recovery process. The results clearly show the superiority of MEWMA over univariate EWMA chart.

Keywords: Multivariate EWMA chart, Normality test, Univariate EWMA chart, Non-normal transformation.

1 Introduction

For more than a decade, there has been increasing interest in monitoring the performance of patients after cardiac surgery, as demonstrated by measurements of certain characteristics related to the well-being of the patient [1,2,3,4]. Quality is seen as important not only because of its potential to detect unacceptable surgical results, but also because of the need to ensure quality when training the next generation of surgeons in a high-risk specialty [5]. All processes including all aspects of medical care are assumed to be subject to intrinsic random (common-cause) variation. The purpose of quality control charts is to distinguish between random variation and special-cause variation, which arise from factors extrinsic to the process of patient recovery. Reducing special-cause variation requires identifying factors that cause the patient's condition to become out of control, and taking appropriate corrective action to improve patient care. A quality control chart can take one of several forms, depending on the type of data which are continuous, binary, or count data. For example, blood pressure and heart rate are continuous data, mortality is binary data, and complications are count data. Shewhart control charts were designed for monitoring batches of results [6]. In the surgical context, a batch might be a series of operations performed over a period of time. Although these charts have been applied in cardiac surgery [4,6], their value for ongoing monitoring of individual results is limited; particularly for multivariate procedures. Another commonly used control chart is the Exponentially Weighted Moving Average (EWMA). This chart has been shown to be most suited for detecting small and persistent process changes [7,8].

Technology has made it relatively easier to record real time data, to keep track of patient performances. This real time performance tracking record in surgical procedures helps doctors to monitor their patient's progress in Intensive Care Unit (ICU) and identify which quality characteristic(s) has high quality variation when compared with their respective specification spread. Control charts are used to assess whether quality characteristic's measurements fall outside the given specification spread. In terms of surgical procedure quality, the possibilities for quality enhancement are major considerations [3,4,9,10].

In this paper, we review the commonly used univariate and multivariate EWMA charts for monitoring the continuous variables in the clinical area. We consider their implementation and performance in light of the challenges faced in heart surgical procedures. The initial focus of this paper is on the use of MEWMA charts to monitor nine non-normal correlated quality characteristics. Attention then shifts to the use of univariate EWMA control charts to identify the variable or group of variables that cause the out-of-control signals in the MEWMA chart.

The paper proceeds as follows: the information of clinical background and data collection is provided in Part 2. We outline the general univariate and multivariate frame work and explain how to construct the MEWMA chart in Part 3. In particular, we address the fact that clinical monitoring often deals with multivariate non-normal correlated characteristics. The methodology of assessing correlation structures between the quality characteristics of interest and transformation of non-normal data are summarized in Part 4.

The discussion is supported by a case study that involved monitoring the medical progress of patients in an Intensive Care Unit (ICU). The paper relates closely to a study carried out at the St. Vincent's Hospital ICU department in Melbourne, Australia. Finally, we provide recommendations on chart selection and implementation in Part 5.

2 Clinical Background

Patients undergoing cardiac surgery including Coronary Artery bypass grafting or Valve replacements are normally placed in the ICU for approximately 12-48 hours for routine monitoring of vital signs includes simple and complex aspects of heart function. This is usually achieved by screening *Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Mean Arterial Pressure (MAP), Systolic Pulmonary Artery Pressure (SPAP), Diastolic Pulmonary Artery Pressure (DPAP), Mean Pulmonary Artery Pressure (MPAP), Pulmonary Capillary Wedge Pressure (PCWP), Central Venous Pressure (CVP) and Heart Rate (HR)*. Most patients appear to progress through this early post-operative phase without serious problems, while others can experience difficulty with blood pressure, sometimes requiring administration of drugs (inotropic drugs) to increase blood pressure. The cardiovascular function and physiological principles of the heart to some extent can be very complicated. In its simplistic state, the heart is a pump; it consists of four chambers (left and right atriums, and left and right ventricles). The right ventricle pumps blood through the lungs for oxygenation, and the left ventricle pumps the oxygenated blood to whole body. To function properly, there must be enough fluid within the main pumping chambers (left and right ventricles).

Some patients in the post-operative care unit suffer from what is called low *Systemic Vascular Resistance (SVR)* syndrome. They appear to be recovering well in the early phase of their post-operative period, but then develop low SVR. In order to maintain blood pressure, the heart has to increase its output (*CO: Cardiac Output*), but is rarely successful. As a result, the *Mean Arterial Pressure* (MAP) begins to fall. This can lead to impaired kidney function and a build-up of acid within the body. Drugs which increase blood pressure may have to be administered, but this can be accompanied by certain risks.

In this paper, an investigation is presented relating to whether the quality characteristics for certain surgical outcomes can be regarded as following the laws of statistical probability. If this is determined to be the case, the opportunity will exist for analysing the data by standard Statistical Quality Control (SQC) techniques to determine when the condition of the individual patients has deviated from the norm, and what this variation implies with regard to SQC. A major objective in this case will be to distinguish this random fluctuation from real changes (favorable or unfavorable) in the surgical results [2,3,10].

2.1 Data Collection

When patients return to ICU, observations on all quality characteristics are documented. The data consist of a series of parameters measured on patients following cardiac surgery. Patients are categorized into three groups based on the values of SVR: Code (1) patients whose SVR is definitely normal (>800), Code (2) patients have mild reductions in SVR (500-800), and Code (3) patients with very low SVR (<500).

For the purpose of this study the authors have decided to concentrate on data of Code (3) patients, since they would be more likely to experience greater variability within the data compared to the norm. Measurements were made half hourly for the first 4 hours and hourly thereafter. For each patient, there are 23 recorded data per characteristics. Twelve quality characteristics were chosen for analysis. In this paper, we only provide the results based on nine correlated characteristics.

3 Methods

3.1 Exponentially Weighted Moving Average (EWMA)chart

The EWMA chart was developed by Roberts in the late fifties and later developed further by other author Hunter [9]. An example of the use of EWMA chart in an intensive care unit can be found in Pilcher [8]. The chart is effective for detecting small to moderate shifts. It is proposed for applications to the process such as chemical industries, financial and management control systems particularly when the sample size is one. The chart can be used to control: an individual sample, the average of a sample, a ratio or a proportion of mortality in the sample.

The EWMA chart is very insensitive to normality assumptions; therefore it is an ideal chart for individual observations in clinical area where most characteristics do not follow a normal distribution. In this chart, the older the observation the less weight it conveys. The plotted statistics are exponentially weighted moving average defined by:

$$z_i = \lambda x_i + (1-\lambda) z_{i-1} \quad , 0 < \lambda \leq 1 , z_0 = \mu_0 \qquad (1)$$

Where λ is a constant and the starting value (required with first sample at $i = 1$) is the process target, In fact z_i is a weighted average of all previous sample means,

$$z_i = \lambda \sum_{j=0}^{i-1} (1-\lambda)^j x_{i-j} + (1-\lambda)^i z_0 \qquad (2)$$

And the variance z_i is

$$\sigma^2_{z_i} = \sigma^2 \left(\frac{\lambda}{2-\lambda}\right)[1-(1-\lambda)^{2i}] \qquad (3)$$

When i increases:$(1-\lambda)^{2i} \to 0$, therefore

$$\sigma^2_{z_i} = \sigma^2\left(\frac{\lambda}{2-\lambda}\right) \quad (4)$$

The control limits for this chart are as follows:

$$\begin{cases} \text{UCL} = \mu_0 + L\sigma\sqrt{\frac{\lambda}{(2-\lambda)}[1-(1-\lambda)^{2i}]} \\ \text{Center line} = \mu_0 \\ \text{LCL} = \mu_0 - L\sigma\sqrt{\frac{\lambda}{(2-\lambda)}[1-(1-\lambda)^{2i}]} \end{cases} \quad (5)$$

3.2 Multivariate Exponentially Weighted Moving Average control chart (MEWMA)

To simultaneously monitoring p correlated quality characteristics, the multivariate EWMA (MEWMA) chart can be applied. Ignoring correlation between variables is the main weakness in the present practice of using independent univariate charts to track each of the p quality characteristics individually.The MEWMA introduced by Lowry et al. (1992) which is a logical extension of the univariate EWMA and is defined as follows:

$$Z_i = \lambda x_i + (1-\lambda)Z_{i-1} \ , Z_0 = 0 \quad (6)$$

Where$\lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$ and $0 \le \lambda \le 1$, and 1 is the identity matrix.

The quantity plotted on the control chart is

$$T_i^2 = Z_i' \sum_{z_i}^{-1} Z_i \quad (7)$$

Where the covariance matrix is

$$\sum_{z_i} = \frac{\lambda}{(2-\lambda)}[1-(1-\lambda)^{2i}]\sum \quad (8)$$

The upper control limit in this paper is produced by the statistical package MINITAB 16 based on $ARL = 200$. Analogous to the situation in the univariate case, the MEWMA chart is equivalent to T^2(or chi-square) chart if $\lambda = 1$. It is possible to place the control limits at a desired level which is called the accepted fluctuation level for a particular quality characteristic. Most cardiac operations are performed by an expert surgical team; they exactly know the outcome of patients and surgical complications such as myocardial infarction, stroke or death.

4 Statistical Analysis

4.1 Normality test and data transformation

The normality test is carried out on all nine quality characteristics for each patient; as an example, the result of this test for patient 10 is presented in Figure 1. It was observed that some characteristics follow a normal distribution with $p-values$ significantly above 0.05. Those characteristics that do not follow a normal distribution have been transformed using inverse, lognormal (ln) and Johnson transformations. The optimal transformation is then selected based on the corresponding $p-value$. We have also compared means, standard deviations, and medians of the individual characteristics among patients using graphical techniques and statistical tests.

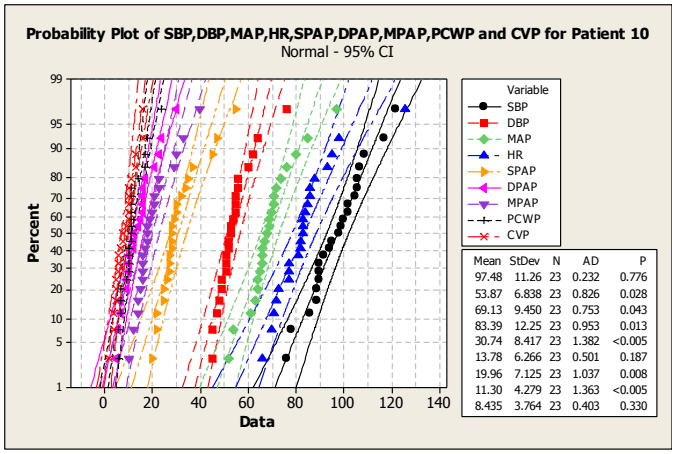


Figure 1: Normal probability plot of 9 characteristics of patient 10

Figure 2 presents the comparison of the DPAP among 12 patients. It is clear that the DPAPs for individual patient have different standard deviation, median, and 25% - 75% percentile values. This pattern has been observed for all characteristics of the study.

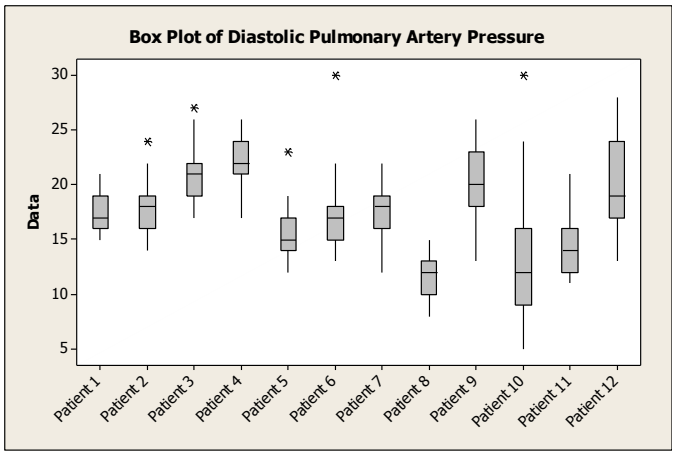


Figure 2: Box plot of comparison of DPAP reading among 12 patients

The results indicate that patients have different summary statistics according to each quality characteristic. As expected, the estimated means of quality characteristics for each patient has been far from the means presented to us by the medical practitioners.Thus, we have used the data for each patient to estimate the overall mean and standard deviation for the individual quality characteristics. The results presented here are based on the estimated means and standard deviations.

4.2 Control Charts

The MEWMA chart is deployed to monitor the performance of each patient based on the nine characteristics that define the well-being of the patient. In this paper, our focus is on patient 10. Through normality test,the quality characteristics for this patient are approximately follow a multivariate normal distribution.

Table 1 indicates the correlation matrix of quality characteristics of patient 10. The highlighted cells show the statistically significant relationship between the characteristics. For instance, the SBP and DBP are correlated with each other (p-value<0.05, r = 0.836); also in practice, these two characteristics together determine the usefulness of blood flow within the body. Hence, a common approach is to use a multivariate procedure to monitor the quality characteristics concurrently.

Table 1: Correlation Matrix of 9 characteristics for patient 10

	SBP	DBP	MBP	HR	SPAP	DPAP	MPAP	PCWP
DBP	0.836							
	0.000							
MAP	0.918	0.975						
	0.000	0.000						
HR	0.023	0.338	0.222					
	0.917	0.115	0.309					
SPAP	-0.241	0.015	-0.090	-0.079				
	0.267	0.945	0.682	0.721				
DPAP	-0.240	0.094	-0.029	0.089	0.918			
	0.270	0.671	0.894	0.687	0.000			
MPAP	-0.294	0.018	-0.103	0.058	0.972	0.965		
	0.173	0.936	0.639	0.794	0.000	0.000		
PCWP	0.023	0.421	0.296	0.633	0.603	0.689	0.698	
	0.916	0.046	0.171	0.001	0.002	0.000	0.000	
CVP	-0.333	0.001	-0.106	0.068	0.860	0.927	0.921	0.672
	0.120	0.998	0.629	0.757	0.000	0.000	0.000	0.000

Figure 3 shows the MEWMA chartof this patient with λ=0.2. According to the plot, if all the nine characteristics are monitored at the same time, the condition of the patient is in control only at reading time 14, 15 and 16 (round points).

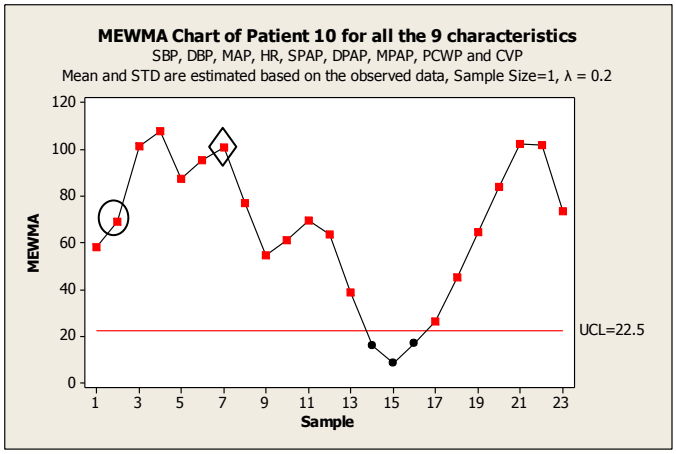


Figure3: Multivariate EWMA chart for patient 10

It is a common practice in statistical quality control to use the individual charts to identify the characteristic(s) responsible for the multivariate out-of-control signals. For this reason, the univariate EWMA chart is imposed on each individual quality characteristic. Out of nine quality characteristics for this patient, five characteristics show the "in control" process. Here, individual charts of other characteristics that detect some out-of-control signals are presented (Figure 4-7).

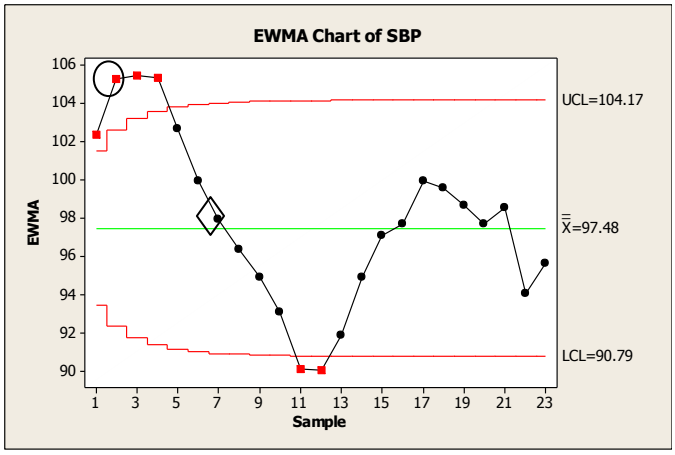


Figure 4: Univariate EWMA chart of SBP for patient 10

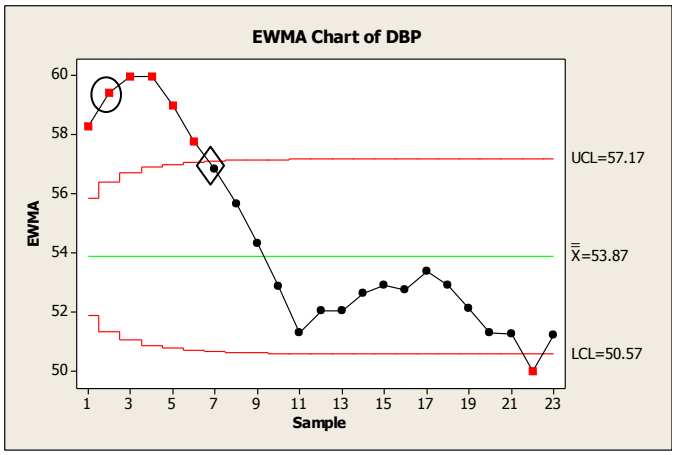


*Figure 5:*Univariate EWMA chart of DBP for patient 10

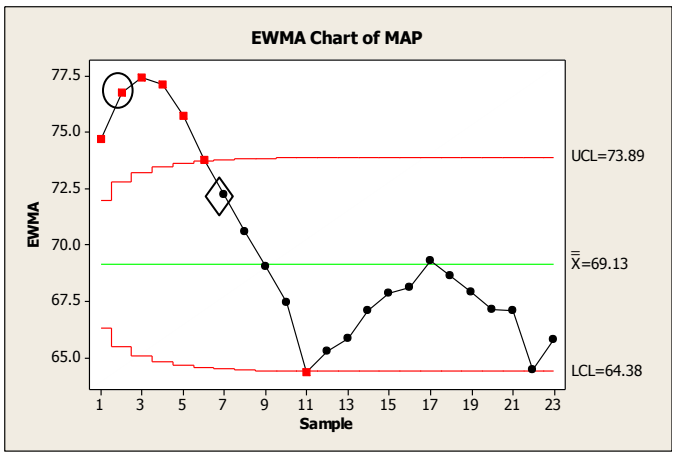


Figure 6: Univariate EWMA chart of MAP for patient 10

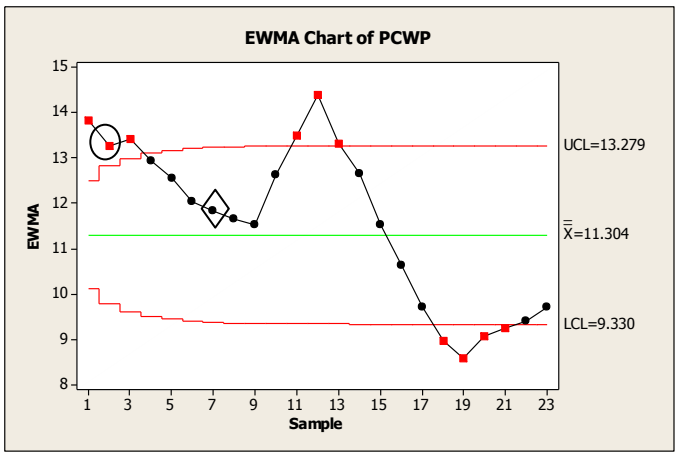


Figure 7: UnivariateEWMA chart of PCWP for patient 10

Based on the plots, once the characteristics are monitored individually, the control charts tend to produce fewer out-of-control signals. However, there are many situations in which the simultaneous monitoring of two or more correlated quality characteristics is necessary.

By considering a multivariate out-of-control signal in MEWMA chart (Figure 3) at reading time 2, the individual charts are examined to identify the source of cause for this alarm at that specific time. The univariate EWMA charts indicate that the four characteristics (SBP, DBP, MAP and PCWP) can be responsible for this alarm, since the out-of-control signals are detected at time 2 in the individual charts. Because there is a significant correlation between these characteristics based on Table 1 (Small $p-value$ is an indication of significant correlation between these characteristics), the multivariate EWMA chart will produce the out-of-control signals more accurately than the univariate EWMA charts.

On the other hand, it is possible that MEWMA chart detects an out-of-control signal, while univariate EWMA charts show in control process. For instance, MEWMA chart (Figure 3) identifies an out-of-control signal at reading time 7, but none of univariate charts (Figure 4-7) show an alarm at this time. The reason would be the correlations between the proposed characteristics (see Table 1) that are not taken into account in univariate charts.

5 Conclusion

In this paper a basis has been presented for improving achievable quality in the multi-variable patient monitoring situation, by applying the multivariate control chart methodology in place of the univariate approach. It has been demonstrated that univariate charts, in some cases, can give a misleading indication that the patient's condition is in control, when the multivariate approach would have properly flagged the problem. It is also shown that when there are nine quality characteristics, which describe the clinical condition of the patients, the individually monitoring would lead to false conclusions. However, we have observed cases where univariate charts signal out of control, while the multivariate chart produces no out-of-control signal. This case may lead to unnecessary treatment of the patient and consequently additional medical costs arise from a "false alarm" in the univariate chart. There are considerable potential benefits to be gained by Intensive Care Units and other surgical areas through the upgrading of quality control procedures by the use of multivariate control techniques.

6 Acknowledgement

The authors would like to thank Dr. John Santamaria, of the Intensive Care Unit, St. Vincent's Hospital, Melbourne, for his useful suggestions and advice during the course of this research activity.

7 References

[1] de Leval, M. R. et al, "Analysis of a cluster of surgical failures", *Journal of Thoracic and Cardiovascular Surgery*, discussion; Vol. 923-4, March 1994.

[2] Noyes, L. "Control charts, Cusum techniques and funnel plots, A review of methods for monitoring performance in healthcare", *Interactive Cardiovascular and Thoracic surgery*, Vol. 9: 494-499,2009.

[3] Rogers, C.A. et al, "Control chart methods for monitoring cardiac surgical performance and their interpretation", *Journal of Thoracic and Cardiovascular Surgery*; Vol. 128: 811-819, Dec 2004.

[4] Sahian D, Williamson W, Svensson L, Restuccia J, D' Agostino R, "Applications of statistical quality control to cardiac surgery", *Ann Thorac Surg*; Vol. 62:1351–1359,1996.

[5] Lawrance R, Dorsch M, Sapsford R, Mackintosh A, Greenwood D, Jackson B, et al. "Use of cumulative mortality data in patients with acute myocardial infarction for early detection of variation in clinical practice: observational study" *BMJ*, Vol. 323:324–327,2001.

[6] Mohammed MA, Cheng KK, Rouse A, Marshall T. "Shewhart's forgotten lessons", *Lancet*, Vol. 357:463–467, 2001.

[7] Montgomery D. C, "Introduction to Statistical Quality Control" (5th Edition) John Wiley & Sons, 2005.

[8] Pilcher, D. V., Hoffman, T., Thomas, C., Ernest, D. & Hart, G. K, "Risk-adjusted continuous outcome monitoring with an EWMAchart: could it have detected excess mortality among intensive care patients at Bundaberg Base Hospital?", *Crit Care Resusc*. Vol. 12: 36–41, 2010.

[9] Hunter, S. J. "The exponentially weighted moving average", *Journal of Quality Technology*; Vol. 18: 203–210, 1986.

[10] Poloniecki, J. et al, "Cumulative risk adjusted mortality chart for detecting changes in death rate: Observed study of heart surgery", *British medical Journal*, Vol. 316;7146: 1-5,1998.

Investigating the Unofficial Factors in Google Ranking

Amir Mardani, Babak Akhgar, Simon Andrews, Simeon Yates [1,] and Mohammad Hassanzadeh[2]
1.Faculty of ACES, Sheffield Hallam University, Sheffield, UK
2.Tarbiat Modares University, Tehran, IRAN
inadram@gmail.com, s.andrews@shu.ac.uk, b.akhgar@shu.ac.uk, hasanzadeh@modares.ac.ir

Abstract

This paper evaluates the effectiveness of some "unofficial" factors in Search Engine Optimisation. A summary of official Google guidelines is given followed by a review of "unofficial" ranking factors as reported by a number of experts in the field of Search Engine Optimisation". These opinions vary and do not always agree. Experiments on keyword density, web page titles and the use of outbound links were conducted to investigate the expert's hypotheses by analysing Google result pages. The results demonstrate that webmasters should avoid having unnecessary outbound links, while attempting to repeat the important keywords of each page one time in their titles, to increase the pages ranking in the results page.

Keywords: SEO, Search Engine Optimisation, SEO unofficial factors

1. Introduction

Every month, more than eighteen billion web searches are performed on the Internet [1]. For companies and individuals have become reliant on the "lower cost", "focus" and simplicity of the Web as a route to market, customers and clients [2] [3]. Therefore sagacious business managers, looking for ways to improve their website's ranking status in Search Engines Result Pages (SERPs) use Search Engine Optimisation methods (SEOs).

In addition to well-known SEOs, based on published factors in Google's ranking process, there are a number of *unofficial* ranking factors that have never been confirmed or denied by Google, that SEOs may exploit. This paper investigates some of these unofficial factors and explores some of the variables involved to thereby recommend appropriate SEOs to exploit them.

Firstly, SEO is explained in relation to official ranking factors published by Google [4]. The research then focuses on unofficial factors which may have an effect on the ranking of a website in search results. The outcomes of this research could be useful for webmasters and site owners who want to augment their viewer density through the Google search engine.

2. A review of SEO factors

There are numerous search engines but only some of them have been successful in attracting large numbers of users [5]. It therefore makes sense for webmasters to implement SEOs that target the most widely used search engines. This segment of the study examines SEO factors pertinent to Google, arguably the most important search engine [6, p. 1].

What is SEO?

SEO is set of small modifications to segments of a website that can assist in getting more hits from search engines [7, p. 1]. There are over two hundred signals that Google considers when ranking websites while scoring their respective search result [8] but Google, in a guidelines for webmasters, officially only cites a limited number of them.

Official factors

Google introduced useful tactics and factors that can help webmasters get a better accessibility status in Search Engine Results Pages (SERP). Google has guidelines relating to page title, site speed, content, anchor text, URLs, navigation, head tags, images and links. Although following these guidelines is certainly effective and can assist search engines to index and crawl websites more easily, they "won't tell you any secrets that'll automatically rank your site first for queries in Google (sorry!)" [4]. A summary of each guideline follows:

Title

Google suggests that webmasters should have unique titles which describe the content of each page accurately [4].

Site speed

"Site speed shows how quickly a website responds to web requests" [9]. Google includes this signal in its search ranking algorithm to encourage webmasters to compact their website [10] .

Content

Creating unique and fresh content for users with relevant information helps Google to reach its goal to "give people the most relevant answers to their queries

as quickly as possible" [11]. Therefore, useful content is one of the most important signals that Google considers in its ranking algorithm. Google uses various criteria to evaluate the quality of the content such as checking the similarity of the content, attractiveness of the topic for the visitors, rationality and comprehensiveness [12].

URL

Google considers the URL of the pages as a signal for ranking websites [13] and asks webmasters to have a descriptive URL for categories and filenames [4].

Navigation

Navigation can help Google to find out important content of each website as well as guiding visitors to find their desired content quickly. Google suggests webmasters plan navigation based on their homepage wisely with a "navigational menu", "text-based links" or a "user-viewable site map" [14].

Anchor text

Anchor text is a clickable text that a user sees on a link [4]. Google asks webmasters to have short but descriptive anchor texts to describe the content and importance of their pages to search engines [15] [4].

Head tags

Webmasters can use concise phrases when describing the content of a page via multiple HTML heading size tags such as "<h1>", "<h2>" and "<h3>". These are important to inform the search engine about the hierarchical structure of the website and the relative importance of text. Although styling the text might achieve the same visual presentation, it does not provide the same meaning or metric to the search engine that a head tag does [8].

Optimise images

Google suggests webmasters put related content around their images and use brief but descriptive text in the "alt" attribute to provide image-related information for their pages. In addition it is quite useful to have a brief but descriptive file name for images rather than generic names such as "pic.gif" or "1.jpg". Google also asks that images be grouped according to size into directories [4] [16] [17] to help Googlebots distinguish the topic of their pages [8].

Link

A website with a proper linking structure can help both Google and users to have better exploration experience and also help it to achieve better visibility in search results [18]. Google uses mature text-matching algorithms to return pages which are both relevant and important for each search query and links are one of the most important factors which can get pages "authority" and "importance". In fact, Google consider a link between pages A to B as a *vote* from A to B and the importance of page A is carried over to page B as "link juice".

On the other hand, Google penalises websites which try to manipulate the search engine by putting unnecessary keywords in their content or copyright content at their end [19]. Google strictly asks webmasters to avoid using keywords excessively in their URLs, Anchor text and images [4]

Google Unofficial factors

Although aligning the website structure and functionality with official factors is good practice, using effective unofficial factors can act as a powerful competitive advantage. Unofficial ranking factors are extensively argued over by SEO experts. Some of these factors are rejected by search engines as cheating, such as "link farming" [20], "clock threading" [21], "hidden text" [22] and "automated queries" [23] but there are other methods that may be effective that are neither officially accepted or rejected by Google. The following sections examine some of these unofficial SEO factors, namely "Best title", "Keyword density" and "outbound link".

Best title

"Do keep it short" says Grappone and Couzin [24, p. 173]. Most search engines present only the first 60 characters of the title in their search result; therefore webmasters should keep their titles short [7, p. 64] [24, p. 173] [25, p. 60] [26, p. 29]. In addition Grappone and Couzin strongly recommended avoiding repeating keywords in titles [24, p. 173]. Similarly Peter Kent believes in short titles but recommends inclusion in the title of the most important keyword of the page [27, p. 35]. However, Konia in "WebPosition Gold", a famous "black hat" SEO tool, recommends webmasters use their primary keywords in the title tag *at least* once. He said webmasters can attract more traffic by using the same keyword in the title multiple times but in different rows. He also stood against the short title idea and suggests webmasters can use longer titles to achieve a better position in search results [28, p. 133]. Enge *et al.* also advocate long titles: "Target longer phrases if they are relevant" [6, p. 212]. Enge *et al.* and Fox believe that having more accurate and descriptive titles are better than simple titles which may be ambiguous or convey less information about the content [6, p. 212] [29, p. 147]. However Google suggests both views have merit, recommending titles that are brief but also descriptive [4].

Keyword density

Keyword density, or in other words the number of times that a specific keyword is repeated in the content, is one of the most important factors that almost all SEO experts believe in. However, there are different points of view about the best keyword density percentage for generating better results.

Jerkovic believes that a good keyword density is between 0.2% and 4%. At the same time he claims that if you go beyond 10%, search engines will penalise you [7, p. 67]. Also, the vendor of WebPosition Gold argues that this percentage could vary from 1% to 4% according to your targeted search engine [28, p. 19]. On the other hand, Kent [27, p. 105] and Baylin [26, p. 135] do not believe that keyword density is a major factor at all. Similarly, Enge *et al*. believe that search engines use more sophisticated analyses than simply counting keywords [6, p. 158]. However, although Google does not encourage webmasters to repeat their keywords within the content of their websites, it has never denied the role of keyword density in SERP.

Outbound links

Outbound links refer to the links which point to external websites. There are webmasters that worry about making outbound links because they think it might cause them to lose their PageRank and also their visitors when they are sending them out of their website. On the other hand, there are some who believe that having only inbound links with no outbound links limits the scope of their website and reduces the quality and richness of the user's experience, and that the best plan is to have a balance between the two [30, p. 268]. Linking to other sites might at first seem ill-advised, in that visitors are being directed away, but it can help visitors find relevant sources. Search engines will find out that you are adding value to the web and improve your site's ranking as a consequence [31, p. 43], particularly when there are links to well-known websites [26, p. 160]. Peter Kent also believes that having good outbound links can help [27, p. 430] while Jerkovic states that having outbound links can actually reduce a website's popularity regardless of the quality of the links. High quality target pages could be considered those having high relevancy or are themselves ranked high. [7, p. 92]. Enge *et al*. believe that having outbound links to mistrusted or poor quality websites can hurt a website's reputation and it's ranking [6, p. 52]. Engaging in so-called linking schemes, where co-operative interlinking of websites is encouraged in an attempt to boost ranking, can back-fire and end up having a negative effect on the ranking.

3. Research Methods

To determine the effect of variables involved in the unofficial SEO methods, an empirical case study on various pre-defined websites was carried out, in a controlled experimental environment. To be confident that rankings were only being affected by the variables under investigation it was important that the other factors were the same in all of the websites. These control factors, such as link structure, site speed and content were ensured by using commercial SEO tools such as *opensiteexplorer.org* and *webseoanalytics.com*.

Population and sample

Data collection came from seocasestudy.co.uk sub-domains which have suitable features to control unwanted factors in SEO experiments. *Seocasestudy.co.uk* is a fresh domain that uses HTML pages for testing SEO approaches in a controlled experimental environment.

Results were collected with Google Custom Search (CSE) [32]. CSE uses the same technology that *Google.com* has and takes into account all the factors which *Google.com* cares about [33].

4. Experimental Findings.

Best Title Experiment

Titles in this experiment are varying in length and keyword repetitions. Three word phrases were created, such as "Top love songs", from which to devise page titles and search terms. Search terms were used consisting of one, two or all three words from the phrase. Page titles were created, given in Table 1, using words X, Y and Z, where X was the first word in the phrase, Y the second and Z the third.

Table 1: Results of the Best Title experiment

ID	Title	Average Rank
1	XYZ	4.8
2	XY	6.9
3	X	8.3
4	XYZ | X	5.4
5	XYZ | XY	4.5
6	XYZ | XYZ	4.5
7	XYZ and XYZ	2.2
8	XYZ | XYZ | XYZ	3.7
9	XYZ | XYZ | XYZ| XYZ	5.2

These nine combinations were tested with various phrases, several times. The same content with the same link structure, and keyword density were published in *seocasestudy.co.uk* subdomains to remove any unwanted factors that might effect the results.

The findings given in Table 1 support Enge *et al*. and Fox, they indicate that having a long title does not harm the rank of web pages. For instance, the titles of the pages in category seven are long in comparison with the search term but still have the best position in the SERP [6, p. 212] [29, p. 147]. The findings also provide evidence that pages that do not have all the search term's keywords in their titles rank lower. High ranks are achieved by having each keyword appear in the title at least one time. The results contradict Grappone and Couzin who argue against having duplicate keywords in the titles [24, p. 173] but support Konia and Kent's idea to repeat the keywords in titles [27, p. 35][28, p. 133]. It also seems that connecting the keywords in the title in a meaningful way could be quite useful. For instance, category seven ranks better than category six by using "and" to give a more meaningful title.

Keyword Density Experiment

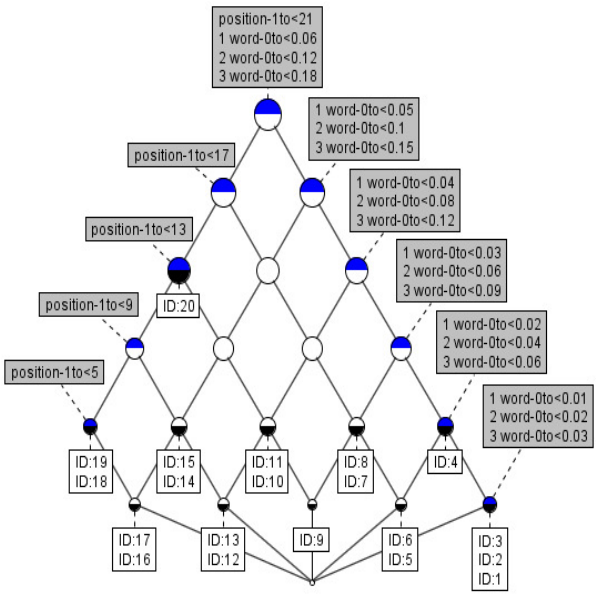
This experiment sought to find the best keyword density to rank better in Google Search engine result page. The experiment was repeated several times for twenty different densities and search terms consisting of one, two or three keywords. For each experiment, the same content and link structure were published in *seocasestudy.co.uk* subdomains to remove any unwanted factors that might effect the results.

Table 2: Results of the Keyword Density experiment

ID	Keyword density			Comparative Rank
	3 word	2 word	1 word	
1	0.0%	0.0%	0.0%	20
2	1.1%	0.8%	0.4%	19
3	2.3%	1.6%	0.8%	18
4	3.4%	2.3%	1.1%	17
5	4.5%	3.0%	1.5%	16
6	5.5%	3.7%	1.8%	15
7	6.4%	4.3%	2.1%	14
8	7.4%	4.9%	2.5%	13
9	8.3%	5.6%	2.8%	11
10	9.3%	6.2%	3.1%	9
11	10.2%	6.8%	3.4%	10
12	10.9%	7.3%	3.6%	8
13	11.7%	7.8%	3.9%	7
14	12.4%	8.3%	4.1%	6
15	13.1%	8.8%	4.4%	5
16	13.9%	9.3%	4.6%	3
17	14.6%	9.7%	4.9%	2
18	15.2%	10.2%	5.1%	1
19	15.9%	10.6%	5.3%	4
20	16.3%	10.9%	5.4%	12

The findings, given in Table 2, do not support the hypotheses put forward within the literature. Jerkovic believes that Google will penalise pages whose keyword densities go beyond 10%. However, using the data from Table 2, a Hässe diagram (Figure 1) was created. In a Hässe diagram, objects (unshaded boxes) are associated with attributes (shaded boxes) that can be reached by traversing upwards from the object. By scaling keyword density against rank, the diagram indicates that the top four out of 20 ranks (IDs 16, 17, 18, 19) had keyword densities between 12 and 18% for a three word search term, between 8 and 12% for a two word search term and between 4 and 6% for a one word search term. The lowest three rankings (IDs 1, 2, 3) had keyword densities of less than 3, 2 and 1%, respectively, for three word, two word and one word search terms; in other words, the higher the density, the higher the ranking. The results also found a linear relationship between keyword density and number of words in the search term. The density ranges used in the scaling were created proportionally to the density values in each successive group of four rankings (ranks 1-4, 4-8, 9-12, 13-16, and 17-20). This revealed a strong proportional relationship between keyword density and number of search term words. This suggests that variation in the number of search term words is not significant in determining ranking and that the function of increase in ranking by increasing keyword density is linear.

Figure 1: Hässe diagram of ranking against keyword density for one, two and three word search terms



Outbound Link Experiment

The experiment sought to find out if pages that have outbound links to high quality content rank better compared to ones which link to low quality content, or have no outbound links at all. In each experiment, the same content and link structure, page title and keyword density were published in *seocasestudy.co.uk* subdomains to remove any unwanted factors that might effect the results.

Table 3 presents the results of the experiment in 9 different groups of pages. The pages in each group were created with the same values, where PR represents the page rank of the target pages and Description is the description of the type of outbound link used to the target page. The PR varies from N/A (has no ranking at all) to 5. To be more precise, pages which are placed in group one have no outbound links at all. Pages in groups 2 to 7 have similar anchor text to the search term whereas pages in group 8 had dissimilar anchor text. Pages in group 9 had indirect outbound links which means that users go through an intermediate page to reach the target page.

Table 3: Results of the Outbound Links experiment

ID	PR	Description	Average Rank
1	n/a	No outbound link	1
2	3	Similar text	6.5
3	n/a	Similar text	6.4
4	4	Similar text	6.4
5	4	Similar text	6.1
6	2	Similar text	6.4
7	5	Similar text	7
8	4	Dissimilar text	2.9
9	4	Indirect text	2.1

Pages which had no or indirect outbound links ranked better in comparison with other pages. The results support Jerkovic's theory that having outbound links reduces the popularity of a webpage regardless of their quality [7, p. 92]. The experiment did not find any strong correlation between having high quality outbound links and getting a better position in SERP.

On the other hand, it seems that pages whose outbound links have similar anchor text to the search term are ranked lower than those with different anchor text. In other words, in searching for "Y", pages which use "X" for their anchor text rank better in comparison with the ones which link to the same page by "Y" anchor text.

5. Conclusion

This paper evaluates some of the unofficial Google website ranking factors put forward by a number of respected SEO experts.

Research findings indicate that titles of the pages should contain search terms at least one time and at the same time results support the idea of repeating keywords in the titles one time to get ranked better. Although findings could not confirm the usefulness of long titles, webpages which had repetitive keywords in their titles did not rank well when compared with others. Small changes in titles, such as connecting keywords with "and" can significantly improve ranking.

Experimental results appear to imply that websites that have no outbound links rank better in comparison with others. However, it could not be confirmed that having high quality outbound links can cause websites to rank better. At the same time, results did not find any strong correlation between low quality links and getting ranked more harshly. In addition, not using keywords within anchor text in outbound links and also using indirect outbound links could be helpful.

The experimental studies found that high ranking can be achieved by having a keyword density of around 5% *per search term keyword*. The function of keyword density against ranking is independent of the number of search term keywords.

In summation of the findings, webmasters should avoid having unnecessary outbound links, while attempting to repeat the important keywords of each page one time in their titles to increase the pages ranking in the results page.

Acknowledgement

The scaling and visualisation techniques used in the analysis of keyword density in this paper are being developed as part of the CUBIST project, http://www.cubist-project.eu/, ("Combining and Uniting Business Intelligence with Semantic Technologies"), funded by the European Commission's 7th Framework Programme of ICT, under topic 4.3: Intelligent Information Management.

6. References

[1] RESTON, VA. comScore Releases January 2011 U.S. Search Engine Rankings. *comscore.* [Online] 2011. [Cited: 8 January 2012.] http://www.comscore.com/Press_Events/Press_Releases/2011/2/comScore_Releases_January_2011_U.S._Search_Engine_Rankings.

[2] SERVE, IB . Advantages and Disadvantages of Internet Advertising. *article alley.* [Online] 2008. [Cited: 7 January 2012.] http://ibserve.articlealley.com/advantages-and-disadvantages-of-internet-advertising-690918.html.

[3] Fron, Christine . Internet Advertising Advantages. *Yahoo! Contributor Network.* [Online] 2005. [Cited: 7 January 2012.] http://voices.yahoo.com/internet-advertising-advantages-1497.html.

[4] Google. Search Engine Optimization starter guide. *Google.* [Online] 2 October 2010. [Cited: 2012 January 7.] http://static.googleusercontent.com/external_content/untrusted_dlcp/www.google.com/en//webmasters/docs/search-engine-optimization-starter-guide.pdf.

[5] Reston, Va. comScore Releases September 2011 U.S. Search Engine Rankings. *ComScore.* [Online] 2011. [Cited: 7 January 2012.] http://www.comscore.com/Press_Events/Press_Releases/2011/10/comScore_Releases_September_2011_U.S._Search_Engine_Rankings.

[6] Enge, Eric , et al. *The art of SEO.* [ed.] Mary Treseler. 1st. Sebastopol : O'Reilly, 2010.

[7] Jerkovic, John . *SEO warrior.* [ed.] Mike Loukides. 1st. Sebastopol : O'Reilly, 2010.

[8] Falls, Brandon , Goradia, Adi and Perez, Charlene . Google's SEO Report Card. *Google Webmaster Central.* [Online] 2010. [Cited: 7 January 2012.] http://static.googleusercontent.com/external_content/untrusted_dlcp/www.google.com/en//webmasters/docs/google-seo-report-card.pdf.

[9] Brutlag, Jake . Speed Matters. *Research Blog.* [Online] 2009. [Cited: 7 January 2012.] http://googleresearch.blogspot.com/2009/06/speed-matters.html.

[10] GoogleWebmasterHelp. Is speed more important than relevance? *Youtube.* [Online] 2010. [Cited: 7 January 2012.] http://www.youtube.com/watch?v=muSIzHurn4U.

[11] Singhal, Amit and Cutts, Matt . Finding more high-quality sites in search. *google blog.* [Online] 2011. [Cited: 7 January 2012.] http://googleblog.blogspot.com/2011/02/finding-more-high-quality-sites-in.html.

[12] Singhal, Amit . More guidance on building high-quality sites. *google webmaster central blog.* [Online] 2011. [Cited: 7 January 2012.] http://googlewebmastercentral.blogspot.com/2011/05/more-guidance-on-building-high-quality.html.

[13] GoogleWebmasterHelp. Does Google consider the URL of an image? *youtube.* [Online] 2009. [Cited: 7 January 2012.] http://www.youtube.com/watch?v=h2SWuUobbr0.

[14] Lee, Jen and Douvas, Alexi . Ring in the new year with accessible content: Website clinic for non-profits. *google webmaster central blog.* [Online] 2010. [Cited: 7 January 2012.] http://googlewebmastercentral.blogspot.com/2010/12/ring-in-new-year-with-accessible.html.

[15] Google. BlogHer 2007: Building your audience. *google webmaster central blog.* [Online] 2007. [Cited: 7 January 2012.] http://googlewebmastercentral.blogspot.com/2007/03/blogher-2007-building-your-audience.html.

[16] —. Image publishing guidelines. *Webmaster Tools Help.* [Online] 2011. [Cited: 7 January 2012.] http://support.google.com/webmasters/bin/answer.py?hl=en&answer=114016.

[17] Linsley, Peter . Get up-to-date on Image Search. *google webmaster central blog.* [Online] 2009. [Cited: 7 January 2012.] http://.blogspot.com/2009/03/get-up-to-date-on-image-search.html.

[18] Szymanski, Kaspar , Far, Pierre and Naumann, Sven . Sharing advice from our London site clinic. *google webmaster central blog.* [Online] 2011. [Cited: 7 January 2012.] http://googlewebmastercentral.blogspot.com/2011/04/sharing-advice-from-our-london-site.html.

[19] Raman. Finding easy-to-read web content. *google blog.* [Online] 2006. [Cited: 7 January 2012.] http://googleblog.blogspot.com/2006/07/finding-easy-to-read-web-content_20.html.

[20] Google. Link schemes. *Google Webmaster Tools Help.* [Online] 2011. [Cited: 7 January 2012.] http://www.google.com/support/webmasters/bin/answer.py?answer=66356.

[21] —. Cloaking, sneaky Javascript redirects, and doorway pages. *Google Webmaster Tools Help.* [Online] 2011. [Cited: 8 January 2012.] http://support.google.com/webmasters/bin/answer.py?hl=en&answer=66355.

[22] —. Hidden text and links. *Google Webmaster Tools Help.* [Online] 2011. [Cited: 8 January 2012.] http://www.google.com/support/webmasters/bin/answer.py?answer=66353.

[23] —. Automated queries. *Google Webmaster Tools Help.* [Online] 2011. [Cited: 8 January 2012.] http://www.google.com/support/webmasters/bin/answer.py?answer=66357.

[24] Grappone, Jennifer and Couzin, Gradiva. *Search engine optimization An hour a day.* [ed.] Pete Gaughan. 3rd. Indianapolis : Willey, 2011.

[25] Michael , Alex and Salter, Ben . *Marketing Through Search Optimization.* 2nd. Oxford : Elsevier Ltd, 2008.

[26] Bailyn, Evan and Bailyn, Bradley. *Outsmarting Google.* [ed.] Sandra Schroeder, et al. 1st. Indianapolis : Que, 2011.

[27] Kent, Peter. *Search Engine Optimization For Dummies.* 3rd. Indianapolis : Wiley, 2008.

[28] Konia, Brad . *Search Engine Optimization with WebPosition Gold.* 2nd. Texas : Wordware, 2002.

[29] Fox, Vanessa . *Marketing in the Age of Google: Your Online Strategy IS Your Business Strategy.* New Jersey : John wiley & Sons, 2010.

[30] Ledford, Jerri . *Search Engine Optimization.* [ed.] Mary Beth Wakefield. 2nd. Indianapolis : Wiley, 2009.

[31] Murray, Glenn. *Seo Secrets.* 2nd. s.l. : Divine Write, 2009.

[34] Google. Google custom search. *Google.* [Online] 28 April 2009. [Cited: 16 January 2012.] http://www.google.com/cse/.

[35] Xu, Hui . is Google costume search engine benefits from the same technology that GOOGLE.com has in ranking. *Google custome search.* [Online] 2011. [Cited: 8 January 2012.] https://groups.google.com/a/googleproductforums.com/forum/#!topic/customsearch/764A29s3stg/discussion.

The Visualization of Collaboration among Iranian Researchers on Nanotechnology: A Social Network Approach

Mohammad Hassanzadeh[1], Reza Khodadust[1], Tahereh Hassanzadeh[2] ,Simeon Yates[3] and Babak Akhgar[3]
1- Tarbiat Modares University, Tehran, IRAN
2- Department of Computer Engineering and IT, Azad University, Branch of Qazvin, Qazvin, IRAN
3- Sheffield Hallam University UK
1.hasanzadeh@modares.ac.ir , reza.khodadust@modares.ac.ir, B.Akhgar@shu.ac.uk 3.t.hassanzadeh@qiau.ac.ir

Abstract— **In this article, we analysed collaboration among Iranian researchers on nanotechnology by using graph-theoretical approaches. Data were collected from Science Citation Index (SCI) via the Web of Science databases advanced search option during 1991-2011 with query made from nanotechnology tree terms, identified by Inspec and Compendex thesaurus. The spatial distribution of network nodes was mapped with the Kamada Kwai algorithm. Spectral clustering algorithm used to identify the clusters for the network of co-authoring countries to analyse the relationship and communities in networks. The period of study divided into 5 sub-periods and performed analysis on each sub-period. Findings well depicted the trend in Iran's nanotechnology research. We also indicated the most productive nodes & nodes with high betweenness centrality, dominant influence, high authoring burst and high novelty from the view-point of intellectual at the author and country level in Iranian nanotechnology research. The primary value of this article lies in extending the understanding of the authoring patterns in Iranian nanotechnology community. This work also has implications for researchers & science policy making.**

Keywords- Iranian nanotechnology; information Visualization; information networks; CiteSpace.

I. INTRODUCTION

To gain insight into today's large data resources, data mining extracts key patterns [16]. Co-authorship studies as kind of data mining approach to bibliographic records provides a window on patterns of collaboration among the academic community [5]. While collaboration is an old concept, the actual study of scientific collaboration is somewhat more recent. At a macro level there are, in general, three major types of factors driving research collaboration: intellectual, economic, and social. At a micro level, collaborative projects are driven by different motivations depending on the specific contexts [14].

In terms of generic network theory, a social network consists of nodes (social actors) and their links (relationships). The collaboration social network in scientific research is a very important category of social networks [15]. Studies use information visualization techniques to enhance comprehension [5]. Some approaches focus on the actors involved and others focus on the set of related activities [14]. Statistical and structural analysis of co-authorship networks can be a useful tool for analysing relationships between various researchers within a scientific field [8]. Such networks reveal the persistent cohesive research collaboration and clustering in the network may represent a knowledge domain [5]. Also, a bibliographic network is important information for researchers when doing a research survey. When a researcher is getting into unfamiliar field of research, grasping overview of the research field, such as bibliographic network, is important [7].

In the twenty-year development vision of Islamic Republic of Iran up to 2025, considerable Emphasis has been placed on scientific development in new technologies, especially in nanotechnology, and in the expansion of scientific collaboration. Such planning and policy making requires the good information concerning available scientific abilities. Although the collaborative research has large benefits in here has been little effort to illustrate international co-authorship and authors' collaboration in the Iranian nanotechnology field. This article is an attempt to fill this gap.

This research aimed to analyse and visualize individual's co-authorship network and the network of co-authoring countries and to measure the rate of co-authorship at the author and country level in Iranian nanotechnology research. The results of this study can inform scientific policy makers on the collaboration level (regional, national and international), the structure of scientific collaboration networks and collaboration patterns.

The remainder of this paper is organized as follows: Section 2 discusses research background. Section 3 describes to data collection and analysis methods. Section 4 presents the collaboration network in the Iranian nanotechnology field. Section 5 provides some concluding remarks.

II. LITERATURE REVIEW

There has been a steady growth in the study of co-authorship in the field of nanotechnology over the last 10 years. In most cases the data for such studies has come from nanoscience journals, nanobank and the ISI Web of Science.

None of these studies provided a mapping the co-authorship network and the network of co-authoring countries.

Schummer [13] carried out analysis of over 600 articles published in "nano journals" in 2002 and 2003, and demonstrated that the average 3.1 nano collaborations had one co-authored article.

Milojevic [11] using social network analysis has traced nanotechnology documents from1970 to 2004. The results indicate that two-thirds of nanotechnology authors are linked together by more than one independent path.

Lin [10] explored the 20 top journals in 2010 in terms of impact factor in nanotechnology counting authors of each document and their collaborators. Characteristics of the distribution of authors in documents and the distribution of collaborators revealed that the distribution of authors is a Weibull distribution and the distribution of collaborators is Poisson distribution.

One only one study, Hassanzadeh and Khodadust [4] explored the linkages between authors in relation to nationality. Hassanzadeh and Khodadust looked at Iranian nanotech research between 1991 to 2010 listed in the ISA Web of Science. They used the search strategy "TS = nano* AND CU = iran" and employed Excel, histcite and pajek software tools. They discovered that Iranian nanotechnology publications from 2 records in 1991 grew to 1,883 records in 2010. More than 80% of articles were published after 2008. The majority of most common co-authors were from Iran (77.8%). Masoud Salavati Niyasr with 133 in the first rank and Alireza Ashrafi with 99 documents in the second rank of the most prolific authors. This paper builds upon this work and explores the visualization of this social-network analysis.

III. METHODOLOGY

Our research focused on all publications of Iranian nanotechnology research between 1991 and 2011. Data were collected on August 5 2011 based on query Made with 65 terms of the nanotechnology tree in the Science Citation Index Expanded (SCI-Expanded) collection via the ISI Web of Science database advanced search.

CiteSpace, a Java application developed by Chaomei Chen of the Information Science and Technology College of Drexel University was used for the co-authorship analysis. In this software there are three sets of threshold levels for co-authorship networks, namely authoring threshold (c), co-authorship threshold (cc), and co-authorship coefficient threshold (ccv). The static visualization of co-authorship network of Iranian nanotechnology publications was done by once choice "author" and threshold 1,2,20; 4,3,20; 3,3,20 and one more "country" and threshold 1,1,1; 1,1,1; 1,1,1. Other options were adjusted as follows: select the 30 top most prolific nodes per slice, and cosine co-authorship coefficient was used for measurement strength of each co-authorship link in within per time slice. The 21-year time interval between 1991 and 2011 was divided into four 5-year time slices (the Beginning from 1991-1995 and the end with 2006-2010) and one 1-year time slice for 2011 (2011-2011).

IV. ANALYSIS OF DATA

4.1. The nature of co-authorship network

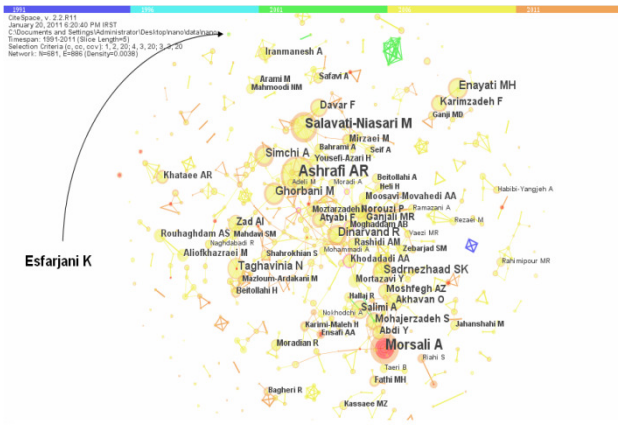
Table 1 summarizes the size of the authoring space and details of individual networks and the merged network. The size of the authoring space in a given time slice is the number of authors that have at least one publication within the given time slice; the size is generally increasing over time. The size for 2011 is smaller as the 2011 data are still incomplete.

TABLE I. TIME SLICING AND THRESHOLD SETTINGS IN CITESPACE FOR CO-AUTHORSHIP NETWORK OF IRANIAN NANOTECHNOLOGY RESEARCHERS

5-year Slices	Criteria			Authoring space size	No. Nodes	No. Links
	C	CC	CCV			
1991-1995	1	2	20	4	4	6
1996-2000	2	2	20	8	0	0
2001-2005	3	2	20	247	22	33
2006-2010	4	3	20	4632	543	630
2011-2011	3	3	20	2501	258	259
Total (Unique)				7392	827(681)	928(896)

Map 1 shows co-authorship network of the most prolific Iranian nanotechnology authors. The co-authoring network in each time slice represents approximately the top 1% prolific authors. The merged co-authorship network consists of 681 unique authors along with 896 co-authorship links among them. Which collectively made 827 appearances in these time slices. In other words, 17.65% of authors appeared in more than one time slice. Publications between 1991 and 1995 are shown in dark blue, 1996 and 2000 in light blue, 2001 and 2005 in green, 2006 and 2010 in yellow and 2011 in grey. Links among authors are co-authorship, which is undirected. The colours of co-authorship links among authors is by the first year rule that two author were co-author. With regard to map 1, the co-authorship network of Iranian Nanotechnology publications consists of the main four super clusters.

The size of each node is proportional to the number of nanotechnology field documents that that node (author) has published. The colours of the tree rings indicate the time patterns of an author in the nanotechnology field. For example, Ashrafi AR Node in middle map 1, the largest authoring circle is filled with colours from yellow to grey. This pattern indicates that Ashrafi AR has published the majority of their documents (65 document) in the fourth five-year period (2006-2010) and that they have not published documents prior to this and has published a very small percentage (ten remaining documents) in 2011. Esfarjani K Node top right corner of map 1 shows a different pattern. Green rings and a thin layer of yellow dominate the Esfarjani K node. This pattern shows that Esfarjani K has published the most of their documents (seven of the eleven documents) in the green time slice; namely 2001-2005.



Map 1. The cluster view from co-authorship network of Iranian Nanotechnology publications. Lines in various colours shows years Where two researchers have published Joint publications. Red circles indicate burst of productivity during the entire interval. Prolific authors have shown with Larger label.

According to Fig. 1, After the Alireza Ashrafi, Ali Morsali and Masoud Salavati Niasari respectively with 63 documents (1.37 Per cent) and 60 documents (1.30 Per cent) allocated 2th and 3th ranks of the prolific authors to themselves. In addition to Mohajerzadeh S were prolific author in terms of production of science in nanotechnology field, has the largest Betweenness Centrality (0.08) toward prolific authors ranked in Fig. 1. Therefore, this author has dominant influence on Iranian nanotechnology field community. Fig. 1 shows the most productive in terms of whole number of documents that they have produced in the entire dataset.

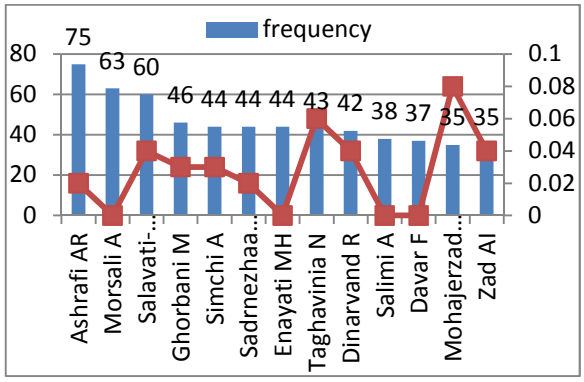


Figure 1. Prolific authors of co-authorship network of Iranian Nanotechnology publications, (There is 6027 citing authors and 17952 co-authorship link between them)

A node centrality will determine the importance of the node's position in a network [1]. Social network analysis provides a set of centrality measures like degree, betweenness, and closeness centrality [9]. In addition to, there are other measures of centrality [6]. A betweenness centrality approach rests on the idea that To what extent may a person control the flow of information due to his or her position in the communication network. The more a person is a go-between. If we consider the geodesics (the shortest path between two vertices) to be the most likely channels for transporting information between actors (the word actor refers to a person, organization, or nation that is involved in a social relation), an actor who is situated on the geodesics between many pairs of vertices is more central [12]. Node is pivotal if its betweenness centrality value were greater and equal to 0.1.

To identify pivotal node, is illustrated only without global pruning version of the larger network. According to table 2, Khodadadi AA, Heravi MM, Golestani-Fard F and Habibzadeh S have respectively The largest betweenness centrality ratio after Moghaddam AB.

TABLE II. PIVOTAL NODES IN THE MERGED CO-AUTHORSHIP NETWORK IN IRANIAN NANOTECHNOLOGY PUBLICATIONS

Row	Author	BC	frequency	Burst	Σ	year	Half life
1	Moghaddam AB	0.15	22	3.21	1.56	2006	0
2	Khodadadi AA	0.12	27	5.61	1.87	2006	0
3	Heravi MM	0.1	11		1	2007	2
4	Golestani-Fard F	0.1	11		1	2006	0
5	Habibzadeh S	0.1	4		1	2007	0

Kleinberg's burst-detection Algorithm can be used to reveal the sharp increases in authoring of one author at nanotechnology field. Burst-detection will determine whether the publication of one author In terms of statistical has considerable surges during short time interval within whole time period. It is valuable for publication analysts to make clear what and when surge the number of a specific author publication. According to table 3, The strongest authoring burst (11.63) has been apparent in Enayati MH authoring history. So that our analysis will demonstrate this author is One of The most active Iranian nanotechnology authors in 2007.

TABLE III. TEN AUTHOR WITH HIGH AUTHORING BURST IN THE MERGED CO-AUTHORSHIP NETWORK OF IRANIAN NANOTECHNOLOGY PUBLICATIONS

Row	Author	Burst	Burst Begin	Burst End	Span	Waiting Time
1	Enayati MH	11.63	2007	2007	1	3
2	Ashrafi AR	11.38	2006	2006	1	0
3	Iranmanesh A	10.61	2007	2007	1	0
4	Karimzadeh F	9.62	2007	2007	1	1
5	Salavati-Niasari M	9.08	2006	2006	1	0
6	Kassaee MZ	8.34	2008	2009	2	0
7	Akhavan O	7.98	2006	2006	1	0
8	Taghavinia N	7.88	2003	2006	4	0
9	Shahrokhian S	7.67	2007	2007	1	0
10	Ganji MD	7.5	2010	2011	2	-1

Sigma (Σ) is introduced as a measure of scientific novelty. Sigma is defined as $(centrality+1)^{burstness}$ [2]. Table 4 shows authors that are likely to represent novel ideas.

TABLE IV. THE NODES WITH HIGH NOVELTY VALUE (> 1) IN CO-AUTHORSHIP NETWORK OF IRANIAN NANOTECHNOLOGY PUBLICATIONS

Row	Author	Σ	year	BC	Burst	frequency
1	Khodadadi AA	1.87	2006	0.12	5.61	27
2	Shahrokhian S	1.84	2007	0.08	7.67	20
3	Mortazavi Y	1.67	2006	0.08	6.81	26
4	Moghaddam AB	1.56	2006	0.15	3.21	22
5	Taghavinia N	1.55	2003	0.06	7.88	43
6	Moshfegh AZ	1.41	2006	0.07	5.21	30
7	Salavati-Niasari M	1.41	2006	0.04	9.08	60

8	Mazaheri M	1.38	2006	0.08	3.98	15
9	Mohajerzadeh S	1.36	2004	0.08	3.92	35
10	Eftekhari A	1.35	2005	0.07	4.19	13

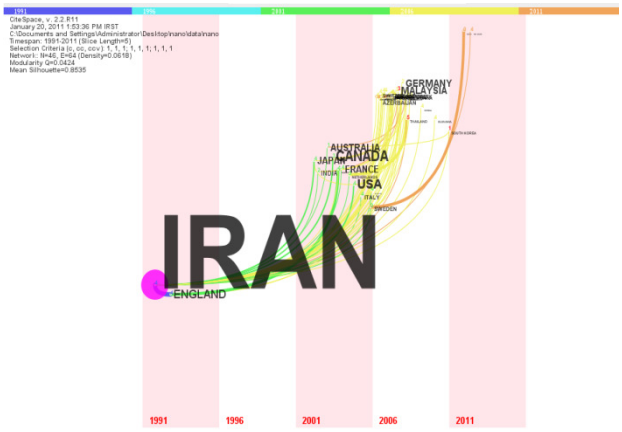
4.2. International Co-authorship

Table 5 shows the size of the authoring space and details of individual networks and the merged network. The merged International Coauthorship network consists of 45 unique countries along with 63 co-authorship links among them. which collectively made 89 appearances in these time slices. In other words, 49.44% of countries appeared in more than one time slice.

TABLE V. TIME SLICING AND THRESHOLD SETTINGS IN CITESPACE FOR THE NETWORK OF CO-AUTHORS' COUNTRIES

5-year slices	Criteria			authoring space size	No. Nodes	No. Links
	C	CC	CCV			
1991-1995	1	1	1	2	2	1
1996-2000	1	1	1	2	2	1
2001-2005	1	1	1	13	13	12
2006-2010	1	1	1	43	43	55
2011-2011	1	1	1	29	29	31
Total(Unique)				89	89(45)	100(63)

with regard to the time zone view, the network of co-authors' countries is consists of the main four super cluster that using spectral clustering algorithm was found these four super cluster, includes seven smaller cluster separate completely.



Map 2. The time-zone view from an un-pruned conceptual modelling of the network of co-authors' countries obtained from Iranian Nanotechnology publications. countries with the highest number of co-authored paper with Iran have shown with larger label.

Fig. 2 shows the frequency of collaboration with different countries in nanotechnology discipline. Iranian nanotechnology collaborated with 44 countries. In nano, the Canada has the highest number of co-authored papers with Iran; America & England is placed respectively after Canada, on the 2nd & 3nd ranks. In addition to Canada were prolific co-author country in terms of collaboration with Iran in nanotechnology field, has the largest Betweenness Centrality (0.09) toward other countries in Figure 2. Therefore, this country has dominant influence on Iranian nanotechnology researchers. In among Asian countries, Malaysia, Japan and India had respectively, the 1nd to 3nd ranks. Taiwan, Ukraine, New Zealand, Morocco, United Arab Emirates, Finland, Oman, Mexico and Libya represent countries that the lowest co-authored paper has been published with Iranian nanotechnology researchers. Iran had a scientific collaboration with nine countries (except Iran) from Organization of Islamic Cooperation (OIC) Fifty-seven Member States during the time period under the study. In terms of the amount collaboration with these countries, Malaysia and United Arabic Emirates had respectively the highest and lowest number of co-authored paper with Iran in SCI during the mentioned years. Azerbaijan, Egypt, Syria, Turkey, Morocco, Libya, Oman is placed respectively after Malaysia, on among the 2nd to 8nd ranks.

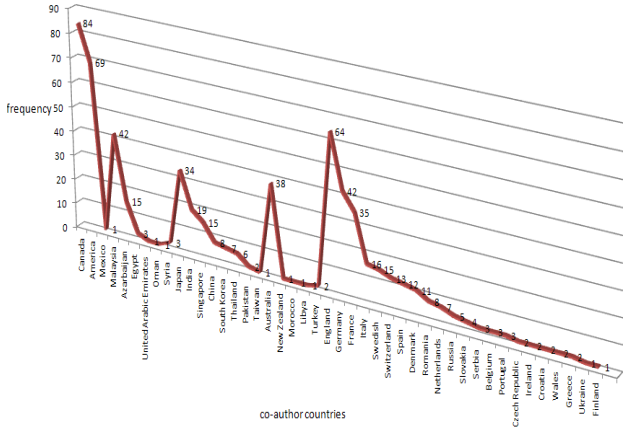


Figure 2. Iranian co-authorship frequency with co-author countries

According to distribution Iranian collaboration with countries of different continents, there are 3 American countries, 14 Asian countries (except Iran), 2 Oceania country, 2 African countries & 23 European countries. Fig 3 demonstrates the situation of co-authored papers of each continent has been published with Iranian nanotechnology researchers by the frequency. As shown in Fig 3, Europe continent had the most number of co-authored papers with Iranian in SCI in comparison with the other continents. Asian & America continents is placed respectively after Europe, on the 2nd & 3nd ranks.

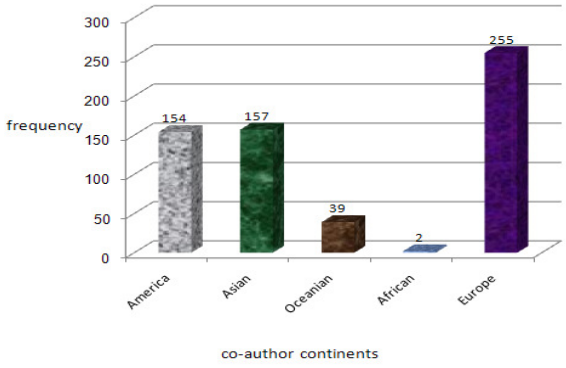


Figure 3. Iranian co-authorship frequency with co-author continents

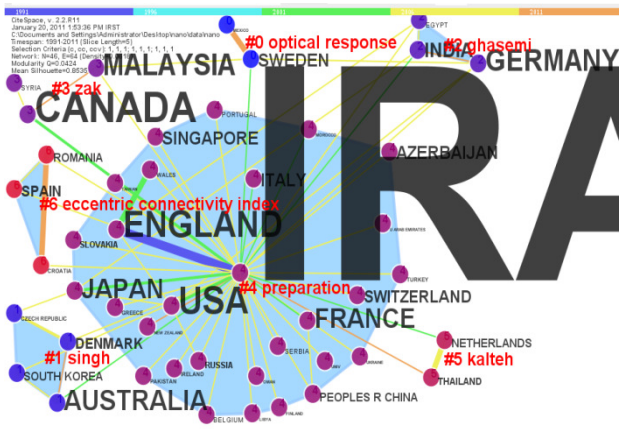
With Sorting an Iran's co-authorship network in the field of nanotechnology with other countries obtained from the Citespce software (Excel file) based on the year, following results are obtained.

TABLE VI. THE BEGINNING CO-AUTHORSHIP OF IRANIAN NANOTECHNOLOGY RESEARCHERS WITH OTHER COUNTRY

First year that iran had Scientific production With other countries	Countries
1991	England
2001	Canada, Australia, Japan
2002	France, India, Netherlands, Wales
2003	America
2004	Italy, Morocco
2005	Swedish
2006	Malaysia, Germany, Azerbaijan, Singapore, Switzerland, Spain, Denmark, Romania, China, Russia, Egypt, Belgium, Portugal, Syria, Turkey, Pakistan, Check Republic, Ireland, Croatia, Greece, Taiwan, Ukraine, United Arab Emirates, Finland, Oman, Libya
2007	Serbia
2008	Thailand, Slovakia
2009	South Korea
2011	New Zealand, Mexico

From 1991 to 2000 (First and second five-year period), the only Colleague or the main colleague of Iranian nanotechnology scientists was from "England". Regardless of Iran can be said that in the third five-year period, the main colleagues were respectively from "Canada", "America", "Australia" and " France" countries. In the fourth five-year period, "Canada", "America", "England" and "Germany" respectively is placed in 1th to 4th ranks. Also, in the end one-year period, Malaysia was Iran's main Colleague.

Capturing the big picture of international collaboration at a macroscopic level and simultaneously linking to subject matters at finer-grained granularities is a long-standing challenge for computational methods [3]. The map 3 is constructed with two layers of information. The base layer is a network of collaborating countries between 1991 and 2011; The thematic layer aggregates individual countries into clusters such that countries in the same cluster have tighter collaboration ties than those in different clusters. The nature of each cluster is characterized by nanotechnology publications collaboratively written by researchers from these countries. Cluster labels shown in map 3 were selected by log-likelihood ratio (LLR) algorithm from the titles of the collaborative publications. In Cluster 4, preparation is the predominant topic for collaborating researchers from Portugal, Singapore, Wales, Taiwan, Morocco, England, Slovakia, Greece, New Zealand, America, Japan, Ireland, Pakistan, Belgium, Oman, Finland, Serbia, China, Italy, Libya, Azerbaijan, Arabic UAE, Iran, France, Ukraine, Switzerland, Turkey and Russia, whereas eccentric connectivity index, in Cluster 6, is likely to be the primary focus of collaborations between Croatia, Spain and Romania.



Map 3. An international collaboration network of 45 countries and 63 collaborative ties. Seven collaboration clusters are identified based on the strengths of collaborative ties.

Countries Betweenness Centrality range is between zero and 1.91. Iran had the largest betweenness centrality. After Iran, respectively, Canada, Sweden, Japan, Denmark and the United States had the highest betweenness centrality. Iran distance range with other countries is between zero and one. Iran has been 41 neighbours. Mexico, Czech-Republic & Syria is not connected with Iran by an independent path namely they is not belong to its domain.

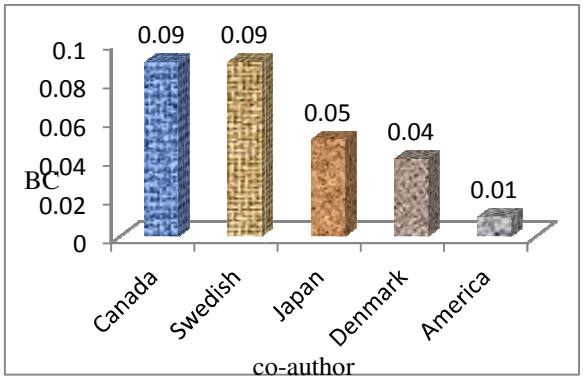


Figure 4. Betweenness Centrality of co-author countries with Iran at nanotechnology

According to table 7, preparation (#4) cluster was the most active cluster in terms of authoring burst and kalteh (#5) clusters is placed on 2th rank. Also, The strongest authoring burst (5.73) has been apparent in France authoring history. So that our analysis will demonstrate preparation (#4) cluster countries were from the most active co-author countries with Iranian nanotechnology researchers to 2007.

TABLE VII. COUNTRIES WITH HIGH AUTHORING BURST IN THE NETWORK OF CO-AUTHORS' COUNTRIES

Row	Author	Burst	Burst Begin	Burst End	Span	Waiting Time	cluster
1	France	5.73	2002	2004	3	0	4
2	Malaysia	5.64	2010	2011	2	4	3
3	America	4.21	2003	2004	2	0	4
4	Thailand	3.45	2008	2011	4	0	5
5	Netherlands	3.44	2008	2011	4	6	5

6	Japan	3.25	2000	2004	5	-1	4
7	South Korea	3.11	2009	2011	3	0	1

According to Figure 5 from the view of Iranian nanotechnology researchers, Japan with a sigma value of 1.16 is allocated the highest rank in terms of to represent novel ideas. Iran's first scientific production with Japan collaboration in nanotechnology field was in 2001.

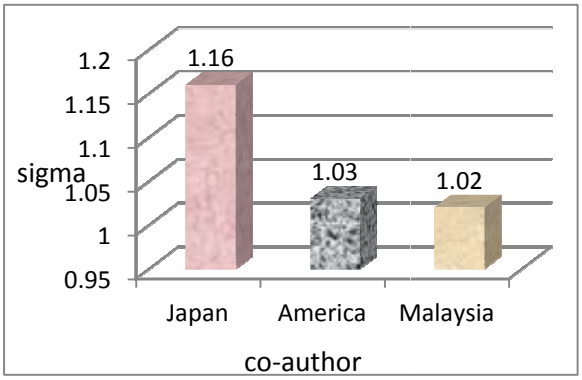


Figure 5. the nodes with high novelty value (> 1) in the network of co-authors' countries

V. CONCLUSION

In this article, we strived to discover the trends & patterns of co-authorship network in Iranian researchers on Nanotechnology. We identified a community of Iranian nanotechnology in terms of the most publishing actors & actors with high betweenness centrality, dominant influence, high authoring burst and high novelty from the view-point of intellectual at the author and country level.

We believe our work contributed to provide new methodology to articulate co-authorship among scientific documents in national and international levels.

REFERENCES

[1] C. Chen, "Citespace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature", JASIST, Vol. 57, No. 3, pp. 359–377, 2006.

[2] C. Chen, F.I. SanJuan, & J. Hou, "The Structure and Dynamics of Co-Citation Clusters: A Multiple-Perspective Co-Citation Analysis", JASIST, pp. 1-33.[print], Forthcoming.

[3] C. Chen, J. Zhang & M. S. Vogeley, "Mapping the Global Impact of Sloan Digital Sky Survey", *IEEE Intelligent Systems*, 24(4), 74-77.

[4] M. Hassanzadeh, R. Khodadust, "Co-authorship and Co-citation in Nanotechnology: a Social Network Approach", international conferences on Webometrics, Informetrics and Scientometrics & twelveth COLLNET Meeting, 7th, 2011.

[5] T. H. Huang & M. L. Huang, "Analysis and Visualization of Co-authorship Networks for Understanding Academic Collaboration and Knowledge Domain of Individual Researchers", Proceedings of the International Conference on Computer Graphics, Imaging and Visualisation (CGIV'06), 2006.

[6] C. Kiss, & M. Bichler, "Identification of Influencers - Measuring Influence in Customer Networks", Decision Support Systems, Vol. 46, No. 1, pp. 233-253, 2008.

[7] T. Kurosawa, Y. Takama, "Predicting Researchers' Future Activities using Visualization System for Co-Authorship Networks', International Conferences on Web Intelligence and Intelligent Agent Technology, 2011

[8] G. LaRowe, R. Ichise & K. B¨orner, "Analysis of Japanese Information Systems Co-authorship Data", International Conference Information Visualization (IV'07), 11th, 2007.

[9] L. Leydesdorff, "betweenness centrality" as an indicator of the "interdisciplinarity" of scientific journals, Journal of the American Society for Information Science and Technology, Vol. 58, No. 9, pp. 1303-1309, 2007.

[10] D. Lin, Distribution of The Coauthorship in Nano Field, ptc/11-09-27, Sep. 2011, http://collnet.cs.bilgi.edu.tr/.

[11] S. Milojevic, Big Science, Nano Science?: Mapping the Evolution and Socio-Cognitive Structure of Nanoscience/nanotechnology Using Mixed Methods, Ph.D. Thesis, University of California, Los Angeles, 2009.

[12] W.D. Nooy, A. Mrvar & V. Batagelj, Exploratory network analysis with pajek, new york, Cambridge university press, 2005.

[13] J. Schummer, "Multidisciplinarity, interdisciplinarity, and patterns of research collaboration in nanoscience and nanotechnology", Scientometrics, Vol. 59, No. 3, pp. 425-465, 2004.

[14] D. Thakur, J. Wang and S. Cozzens, "What does International Co-authorship Measure?", Atlanta Conference on Science and Innovation Policy, 2011

[15] S. Wen-jun , J. Ai-xian, "The Collaboration Network in China's Management Science", International Conference on Management Science & Engineering, 16th, 2009.

[16] Q. Ye, B. Wu, B. Wang, "Visual Analysis of a Co-authorship Network and its Underlying Structure", International Conference on Fuzzy Systems and Knowledge Discovery, Fifth, 2008.

Implementing Boolean Matching Rules in an Entity Resolution System using XML Scripts

Yinle Zhou*
yxzhou@ualr.edu
Information Science Department, University of Arkansas at Little Rock

John R. Talburt
jrtalburt@ualr.edu
Information Science Department, University of Arkansas at Little Rock

Fumiko Kobayashi
fxkobayashi@ualr.edu
Information Science Department, University of Arkansas at Little Rock

Eric D. Nelson
ednelson@ualr.edu
Acxiom Corporation

Abstract - *This paper describes a method for allowing users to define Boolean matching rules for entity resolution systems using XML scripts. The design allows users the flexibility to create and modify matching rules that are interpreted at run-time rather than having to go through the process of modifying and testing the underlying codebase. The design has been validated through implementation in open source entity resolution system named OYSTER. The paper also describes proposed design enhancement that address the issues of cross-attribute comparison and conflict rules.*

Keywords: Boolean matching rules, entity resolution, OYSTER, cross-attribute comparison, conflict rules

1 Background

ER is the process of determining whether two references to real-world objects in an information system are referring to the same object, or to different objects [1]. Although there are a variety of algorithms for implementing ER processes, at a basic level they all involve making decisions about the similarity of the references. ER systems assert their decisions by appending an identifier value, called a "link" to each record representing a reference. If the decision is that two records are referring to the same entity (i.e. equivalent references), they are given the same link value, else they are given different link values.

Many ER processes still follow the Fellegi-Sunter Model for record linking [2] in which pairs of records are judged to be "link" or "non-link" pairs (or perhaps "possible link" pairs) depending upon which attribute values agree or disagree. For example, the pattern that two student enrollment records agree on first name values, agree on last name values, agree date-of-birth values, but disagree on school identifier values, might be designated a link rule, i.e. the decision is that the records are for the same student. Matching rules that give a yes or no (true or false) decision are called Boolean matching rules [3]. There is another family of entity resolution rules that compute a numerical score representing the distance between the records. However, the design presented here only addresses the implementation of Boolean matching rules.

2 Boolean Matching Rules and Similarity Functions

The implementation of Boolean matching rules generally extends the definition of agreement to allow for some level of variance between the attribute values. In other words, two attribute values might still be considered to agree even if the two values are not identical, but are still in some way similar. The similarity functions [4] that make these determinations have many different forms. For example, some rules might allow common nicknames, such as "Jim" for "James", or common misspellings or confusions, such as, "Johnson" for "Johnston" to be within the threshold of similarity. Boolean matching rules that

allow such variation are sometimes called "fuzzy" matching rules.

Similarity functions for values represented by character strings generally fall into three categories, syntactic, semantic, and phonetic similarity. Syntactic similarity functions, also called approximate string matching (ASM) algorithms [5], measure the degree to which two strings share the same characters in the same order. There are many different ASM algorithms each focusing on a particular aspect of string similarity. For example, one of the most commonly used ASM algorithms is the Levenshtein Edit Distance algorithm [6]. It measures the similarity between two strings in terms of the minimum number of character deletions, insertions, or substitutions that must be performed in order transform one string into another. Using these manipulations, the strings "JIM" can be transformed into the string "JAMES" in 3 operations by starting with "JIM", substituting "A" for "I" giving "JAM", then inserting "E" and "S" giving "JAMES".

In the case of Boolean matching rules, the similarity functions are normalized so that they return a rating value between zero and one where a rating value of one represents an exact match and rating values less than one proportionally lesser degrees of matching. For example the Levenshtein Edit Distance ASM can be normalized [7] as (1)

$$Levenshtein\ Mormalized\ \ Rating = 1 - \frac{Levenshtein\ \text{Distance}}{\max\{\text{Length (String1), Length(String2)}\}} \quad (1)$$

In the worst case, there are no shared characters between String 1 and String 2. If String 1 is longer than String 2, then the extra characters in String 1 must be deleted until String1 is the same length as String 2. Next every one of the remaining characters in String 1 must be replaced by the corresponding character in String 2. The total number of operations is the same as the length of String 1, the longer string. In this case the Normalized Rating value given by the formula is zero. In the case that the two strings are identical, the Levenshtein Distance is zero and the Normalized Rating will be one.

Other ASM algorithms include Jaro [8], Jaro-Winkler [9], Jaccard Coefficient [10], Cosine Similarity [11], q-Gram [12], q-Gram Tetrahedral Ratio [13], and many others. Descriptions, pseudo-code, and even coded modules for these algorithms are available from a number of sources that are easily found through an Internet search.

Semantic similarity functions are based on the meaning of the attribute value rather than syntactic similarity. Perhaps the most common are based on name variants (so called "nicknames") used in a particular language or country. As mentioned earlier "JIM" and "JAMES" are considered semantically equivalent English names even though their Levenshtein Normalized Rating for similarity is only 0.40 or 40%. Semantic similarity functions are usually implemented as a lookup algorithm using predefined tables of equivalent names.

Phonetic similarity functions are based on the similarity of pronunciation of words in a particular language or country, and like semantic similarity it is primarily used for name matching. For example in English, the names "KAREN" and "CARYN" are phonetically identical even though they are semantically different and only have a Levenshtein similarity of 0.60. Phonetic similarity functions often take the form of "hashing" functions that operate on the characters of the input string and create a new output string in which similar sounding characters or groups of characters are assigned the same symbol. The objective is to have two phonetically similar strings transform into the same output string called a "hash token". Perhaps the most commonly used phonetic similarity function for English is the Soundex algorithm [14]. In the Soundex algorithm, the first letter is retained, but vowels are removed and groups of similar sounding consonants are replaced with the same numeric digit. For example in the Soundex algorithm, the names "PHILIP" and "PHILLIP" both create the hash token "P410".

3 Problem Statement

The designs and methods for implementing of Boolean Matching Rules vary widely across ER systems. Three of the most common problems encountered are

Problem 1: The rules are embedded in the application code. This usually means that any rule modifications must go through a sometimes lengthy software change process.

Problem 2: Even if the rules are parameter driven, the semantic encoding of the rule parameters often uses an application specific syntax.

Problem 3: Users may only have a limited number of choices for similarity functions.

4 A Hierarchical Design for Boolean Matching Rules

Figure 1 shows a hierarchical design for Boolean matching rules. Although there may be certain types of rules that do not fit this pattern, the design given here does accommodate rules based on the Fellegi-Sunter agree/disagreement patterns [2] used in most ER systems. The Future Work section of this paper also proposes some extensions to the basic design that will allow it to address an ever broader spectrum of rule patterns.

In the design shown in Figure 1, each T_j is a Rule Term that represents a logical proposition that two attribute values are

similar (true) or not similar (false) as defined by a specific similarity function and threshold. A Boolean matching rule may have several terms, but must have at least one term. An example of a rule term would be that the first names on two student records must be at least 0.80 (80%) similar as measured by the Levenshtein Normalized Rating described earlier in the paper.

In the hierarchical design, a Boolean Matching Rule is a collection of one or more Rule Terms connected with AND logic between the terms, i.e. all rule terms must be true in order for the rule itself to be true. In addition, an ER system may use more than one rule, and when it does, the design asserts OR logic between rules, i.e. the final resolution decision between two records is true if and only if at least one rule is true.

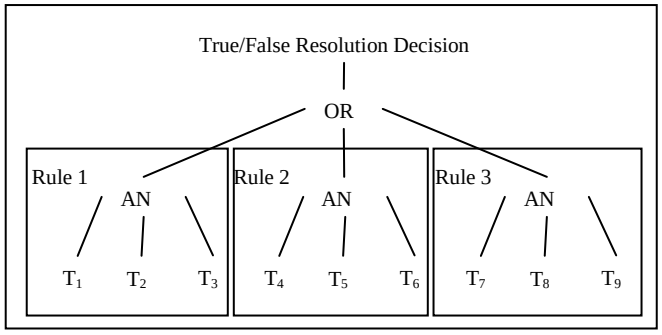


Figure 1: Hierarchical Structure of Boolean Matching Rules

5 An XML Implementation of the Design in OYSTER

The hierarchical design has been validated through its successful implementation as an XML script in an open source entity resolution system called OYSTER [15]. The OYSTER project is the outgrowth of a project at the ERIQ Research Center at the University of Arkansas at Little Rock to create an ER system for class-room support of ER education. Originally introduced in the ER textbook Entity Resolution and Information Quality [1], the Java source code and documentation for OYSTER are now available through the SourceForge.net website indexed under the keyword "oysterer".

The implementation of the design in OYSTER can best be explained by discussing the example XML script shown in Figure 2. OYSTER uses a number of XML scripts that determine its configuration at run-time. The script shown in Figure 2 is an example of the Attribute Script. The attribute script has two major functions in OYSTER. The first is that it defines the attribute name space for a run. The valid labels for OYSTER attributes are defined by a series of <Attribute> elements. Only the labels are defined in the Attribute Script. Other details about the attributes, such as their physical name in a database table or their ordinal position in an input record, are described in a separate Source Descriptor script not shown here.

The rules themselves are defined in the <IdentityRules> Section of the attribute script as a series of <Rule> elements. The <Rule> elements enclose a series of <Term> elements that define each rule term. The value of the "Item" attribute in a <Term> element must be one of the OYSTER attribute labels defined in the previous section of the same script. The label given in the <Term> element designates the OYSTER attribute whose values are to be compared.

The value of the "MatchResult" attribute in a <Term> element designates the similarity function to be used in comparing the two values of the OYSTER attribute given as the value of the "Item" attribute in same <Term> element. In addition, the value of the "MatchResult" also specifies the threshold of similarity that must be met for the term to be true. For example, Rule 3 (Ident="3") defines three terms. The first term requires that the two values of the "First" attribute (student's first name) must be similar at the 0.80 rating or greater using the normalized Levenshtein Edit Distance (LED). The second term requires that the values of the "Last" attribute (student's last name) must be identical strings ("Exact"). The third term requires that the two values of the "SSN" attribute (student's 9-digit social security number) must differ by one transposition of two adjacent digits ("Transpose"), a common data entry error.

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- Document: Attribute.xml, Created on: 3/10/2012, Author: Yinle
Zhou-- >
<OysterAttributes System="Demo">
   <Attribute Item="First" Algo=""/>
   <Attribute Item="Last" Algo=""/>
   <Attribute Item="PNbr" Algo=""/>
   <Attribute Item="SSN" Algo=""/>
   <Attribute Item="DOB" Algo=""/>
   <IdentityRules>
      <Rule Ident="1">
         <Term Item="First" MatchResult="Exact"/>
         <Term Item="Last" MatchResult="Exact"/>
         <Term Item="SSN" MatchResult="Exact"/></Rule>
      <Rule Ident="2">
         <Term Item="First" MatchResult="Nickname"/>
         <Term Item="Last" MatchResult="Exact"/>
         <Term Item="SSN" MatchResult="Exact"/></Rule>
      <Rule Ident="3">
         <Term Item="First" MatchResult="LED(0.80)"/>
         <Term Item="Last" MatchResult="Exact"/>
         <Term Item="SSN" MatchResult="Transpose"/></Rule>
      <Rule Ident="4">
         <Term Item="First" MatchResult="Exact"/>
         <Term Item="Last" MatchResult="Exact"/>
         <Term Item="SSN" MatchResult="Exact"/></Rule>
      <Rule Ident="5">
         <Term Item="First" MatchResult="Soundex"/>
         <Term Item="Last" MatchResult="Exact"/>
         <Term Item="SSN" MatchResult="Exact"/></Rule>
   </IdentityRules>
</OysterAttributes>
```

Figure 2: Example Attribute Script

The following is a list of the similarity functions currently implemented in OYSTER Version 3.2.

Exact – The two values being compared must be identical strings character-by-character.

Transpose – Requires a difference of 2 adjacent characters in one of the strings being in reverse order. Example: Value1 = "12**34**5" and Value2 = "12**43**5".

Missing – Two values being compared meet this matching condition if one or both values are null or blank.

INITIAL – Two values being compared meet this matching condition if one value is a single character and it matches the first character of the other value. Example: Value1 = "J" and Value2 = "John".

NICKNAME – This matching condition is met if both values correspond to one another in the user provided lookup table (the "alias.dat" file). Example: Value1 = "Robert" and Value2 = "Bob".

SOUNDEX –This matching condition is met if the two values generate the same Soundex code. The Soundex algorithm is a phonetic similarity function for names in English. Example: Both "Robert" and "Rupert" return the same Soundex code.

DMSOUNDEX – The Daitch-Mokotoff Soundex Algorithm, a refinement of the Russell and American Soundex algorithms designed to allow greater accuracy in matching of Slavic and Yiddish surnames with similar pronunciation but differences in spelling. Example: Both "Moskowitz" and "Moskovitz" generate the same DMSOUNDEX.

IBMALPHACODE – An implementation of the IBM Alpha Code algorithm. Example: Using this algorithm, both "Rodgers" and "Rogers" produce the same IBM Alpha Code.

NYSIIS – An implementation of the New York State Identification and Intelligence System coding algorithm, a phonetic algorithm devised in 1970 as part of the New York State Identification and Intelligence System. Example: Both "McKee" and "Mackie" produce the same NYSIIS code.

MATCHRATING – An implementation of the Western Airlines Match Rating Approach algorithm. Example: "Byrne" and "Boern" produce the same match rating in this algorithm.

LED(threshold) – The Normalized Levenshtein Edit Distance. If no threshold is specified, the default threshold is 0.80

QTR(threshold) – The Q-Gram Tetrahedral Ratio. If no threshold is specified, the default threshold is 0.25.

SUBSTRLEFT(length) – If N is the length value given, this similarity function requires an exact match on the left-most N characters of both strings. Example: If N=3, then "Samuel" matches "Sam".

SUBSTRRIGHT(length) – If N is the length value given, this similarity function requires an exact match on the right-most N characters of both strings. Example: N=4, then "JeanAnne" matches "Anne".

SUBSTRMID(start, length) – If N is the length value and S is the start value given, this similarity function requires an exact match on the N characters of both strings starting at position S. Example: If N=6 and S=2, then "Krystal" matches "Crystalline".

In addition to the functions listed here, users can create new similarity functions by extending the Java class OysterComparator.java. OysterComparator.java is the base class for all of the similarity functions defined in OYSTER and implements the "Exact" and "Missing" similarity functions.

6 Conclusion

The hierarchical design for Boolean matching rules has been successfully implemented in the OYSTER open source ER system using XML. The implementation addresses all three of the problems given earlier.

Problem 1: Rules embedded in the application code. The implementation described here allows the user to define any number of hierarchical matching rules through simple XML scripting. The matching rules are interpreted at run-time by the system, and rule modifications are accomplished by simply changes the rule definitions in the script.

Problem 2: The semantic encoding of the rule parameters uses an application specific syntax. XML is an open, and well-recognized, standard for scripting. Even though the semantics are application specific, e.g. definition of "LED(0.80)", its encoding as a matching rule follows the standard XML syntax.

Problem 3: Users may only have a limited number of choices for similarity functions. By placing the matching rule definitions in a run-time XML script, the definition of rule is de-coupled from the definitions (coding) of the similarity functions the rule invokes. This allows any number of similarity functions to be separately coded and tested, then invoked as needed through the XML script.

7 Future Work

OYSTER users have requested several matching rule features that were not addressed in the original hierarchical

design. Two of these are cross-attribute comparison and conflict rules.

Cross-attribute comparison is the ability to compare the values in two different attributes. The current design assumes that the similarity functions in a rule term will always compare values for the same attribute. However in many situations, the information in records can be "misfielded". This often occurs when recording names, for example, entering a student's first name in the last name field and the student's last name in the first name field. In order for a matching rule to address this situation the rule semantics and syntax must allow a way for the similarity function to compare values from different attributes

Conflict rules are Boolean matching rules that are applied to a groups of records rather than a single pair of records. Conflict rules help ER systems avoid making false positive errors that can arise through transitive resolution. In order to give consistent results [16], ER systems must support transitive resolution. The problem with pair-wise Boolean matching rules is that even through Record 1 matches Record 2, and Record 2 matches Record 3, it does not always follow that Record 1 matches Record 3. Nevertheless, under this scenario, a consistent ER system would still link Records 1, 2, and 3 together as representing the same entity.

Although this is generally desirable, it can sometimes lead to false positive errors. Suppose that Record 1 and Record 3 are processed first. Because in the scenario they do not match, they would form different identity groups. Then suppose that Record 2 is now processed. Given that Record 2 matches both Record 1 and Record 3, transitive resolution would normally require the ER system to link all three records into a single identity group. That is unless there was a logical conflict between Record 1 and Record 3.

For example suppose that Record 1 represents a female student (i.e. it has a gender attribute value indicated female) and Record 3 represents a male student (i.e. it has a gender attribute value indicating male). Because a system can use multiple matching rules, it is possible that one of the rules allows Record 2 to match both Records 1 and 3 without using the gender attribute, i.e. Record 2's gender attribute may have a missing value, but still matches to Records 1 and 3 using other attributes. A gender conflict rule would prevent Records 1 and 3 from being consolidated into a single group with Record 2. In other words, a conflict rule would override matching rule logic if the application of the rule would join two groups of records with conflicting values, i.e. a record in the first group and a record in the second group disagree on one or more attribute values such as gender or age.

8 Acknowledgement

The research described in this paper has been supported in part by research grants from the Arkansas Department Education and the National Institutes of Health Clinical and Translational Science Awards program.

9 References

[1] John Talburt. "Entity Resolution and Information Quality". Morgan Kaufmann, 2010.

[2] Iven Fellegi and Alan Sunter. "A Theory for Record Linkage"; Journal of the American Statistical Association, Vol. 64 No. 328, 1183-1210, 1969

[3] Steven Whang and Hector Garcia-Molina. "Entity Resolution with Evolving Rules"; Proceedings of the VLDB Endowment, Vol. 3 Issue 1-2, 1326-1337, September 2010

[4] Felix Naumann and Melanie Herschel. "An Introduction to Duplicate Detection". Morgan & Claypool, 2010

[5] Gonzalo Navarro. "A Guided Tour to Approximate String Matching"; ACM computing Surveys, Vol. 33, Issue 1, 31-88, 2001

[6] Vladimir Levenshtein. "Binary Codes capable of Correcting Deletions, Insertions and Reversals"; Soviet Physics Doklady, Vol.10, Issue 8, 707-710, 1966

[7] Peter Christen. "A Comparison of Personal Name Matching: Techniques and Practical Issues"; Sixth IEEE International Conference on Data Mining-Workshops (ICDMW'06), 290-294, 2006

[8] Matthew Jaro. "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida"; Journal of the American Statistical Association, Vol.84, Issue 406, 414-420, 1989

[9] William Winkler. "The State of Record Linkage and Current Research Problems"; Statistical Research Division, U.S. Census Bureau, 1999

[10] Alvaro Monge and Charles Elkan. "The Field Matching Problem: Algorithms and Applications"; In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 267-270, 1996

[11] Pang-Ning Tan, Michael Stenbach, and Vipin Kumar. "Introduction to Data Mining". Addison-Wesley, 2005

[12] Erkki Sutinen and Jorma Tarhio. "On using q-gram locations in approximate string matching"; In Proceddings

of Third Annual European Symposium on Algorthims, 327-340, 1995

[13] Greg Holland and John Talburt. "q-Gram Tetrahedral Ratio (qTR) for Approximate String Matching"; 2010 Annual Acxiom Laboratory for Applied Research Conference (ALAR-10), 2010

[14] Donald Knuth. "Sorting and Serching"; Art of Computer Programming, Vol.3, 391-392, 1989

[15] Yinle Zhou, John Talburt, Ying Su, and Ling Yin. "OYSTER: A Tool for Entity Resolution in Health Information Exchange"; In The 5th International Conference on the Cooperation and Promotion of Information Resources in Science and Technology (COINFO"10), 356-362, 2010

[16] Omar Benjelloun, Hector Garcia-molina, Hideki Kawai, Tait Larson, David Menestrina, Qi Su, Sutthipong Thavisomboon, Jennifer Widom. "Generic Entity Resolution in the SERF Project"; IEEE Data Engineering Bulletin, 13-20, 2006

A Review of Relationship Resolution: Terminology and Classification

Fumiko Kobayashi, John R. Talburt
Information Science Department
University of Arkansas at Little Rock
2801 South University AVE. EIT 550
Little Rock, AR, USA

Abstract- *Traditionally entity resolution (ER) has been based on direct record-to-record matching. However there is a growing body of ER research that focuses on the resolution of two entity references based on their patterns of relationships or associations with other references. ER based on these new methods for performing ER on patterns of association is called Relationship Resolution. As often happens in a new area of research, many papers describe the same concepts using different terms while in other cases, fundamentally different concepts and approaches are described using the same terms. This paper attempts to clarify some of the terminology being used to describe Relationship Resolution and to categorize the Relationship Resolution approaches and algorithms found in the literature.*

Keywords- Relationship Resolution; Entity Resolution; Reference Graphs; Entity Relationship Graphs; Graph Clustering; Record Linkage

1 Introduction

Relationship Resolution is Entity Resolution (ER) that attempts to determine if two references to real-world objects are references to the same or different objects by examining the relationships that the two references have with other references, and not just their similarity to each other (pair-wise record matching). A classic example is the authorship ambiguity problem. In bibliographic citations, first names are often reduced to initials creating a situation in which different authors may be cited using the same name. However, given a number of citations it may be possible to disambiguate between different authors by observing their relationships with co-authors [1] as seen in Fig 1.

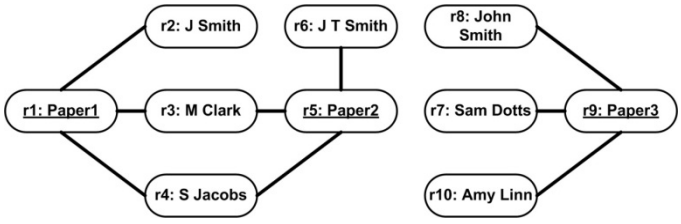


Figure 1. This reference graph represents the relationship between authors (r2, r3, r4, r6, r7, r8, r10) and papers (r1, r5, r9). The authors and papers make up the vertices of the graph while the edges represent the relationship between papers and authors defining which authors wrote which papers.

In Fig 1, if standard pair-wise ER was performed on the attributes in the vertices, then by using an "Initial" similarity function for the first name and an "Exact" function on the last name, the ER system would result in five author entities and three paper entities. The five author entities would be: {r2, r6, r8}, {r3}, {r4}, {r7}, {r10}. This pair-wise matching could result in a false positive if r2 and r6 actually represent "James T Smith" and r8 refers to "John Smith". Relationship Resolution, however, can correct these types of false positive matches by taking the information carried in the edges into consideration. Where standard pair-wise ER merged r2, r6, and r8, a Relationship Resolution algorithm would have taken into consideration their relationships with the other authors. By doing so, it can infer that since r8 does not share any relationships with r2 and r6, then r8 must refer to a different entity. A Relationship Resolution algorithm could use the additional information to infer that there are actually six author entities and three paper entities: {r2, r6}, {r3}, {r4}, {r7}, {r8}, {r10}, and {r1}, {r5}, {r9}.

There are many proposed methods for performing Relationship Resolution for many different types of data. The basic problem is that information comes from multiple sources and in varying formats with no unique identifier that is common across all sources. The information received from some sources may be flat single-typed data which contains only attributes that describe a single entity, or multi-type data in which the attributes in the data refer to multiple types of entities that are related.

With single-type data, ER can be performed through traditional methods which rely on the degree of similarity between the attributes describing the entity such as name, address, and age. These approaches often use approximate string matching algorithms (ASM) [2] or other similarity functions [3, 4] as a way to measure pair-wise similarity. In most cases, these traditional record matching methods do not leverage the potential richness of information based on relationships with other entities.

There are many Relationship Resolution algorithms that have been proposed that provide a method to cluster and resolve entities that are found in multi-type data. In some cases it can also be beneficial to apply Relationship Resolution to single-type data if the relations in the ER graph are analyzed and blended with knowledge about the similarity of the entity characteristics.

2 Problem Definition

With all the new methods that are being proposed for Relationship Resolution there are cases in which the same concepts are being described with different terminology. This paper seeks to clarify the terminology used in Relationship Resolution. The second issue this paper attempts to address is the categorization of the various Relationship Resolution algorithms from different viewpoints. The categorization is based on how the methods perform the clustering and resolution, the type of data they are suited for, or if the algorithms are domain specific.

3 Terminology

Many papers refer to the same concepts using different terminology. For the purposes of this paper, each concept is provided a standardized name, or term, to help clarify the underlying concept and to clarify the meaning of the different terms that are being used in the literature pertaining to Relationship Resolution. Some of these terms are already considered the basic vocabulary for standard ER and others are borrowed from other disciplines. By clarifying the concepts, it will minimize the time and maximize the understanding when associating a new term to an existing concept.

In [5], many of the terms and concepts used in non-relationship ER are discussed and defined. These same concepts can be applied as some of the base vocabulary for Relationship Resolution which is identified in [5] as one of the four methods of ER. The most general concept in entity resolution is *Entity Resolution*, also referred to as disambiguation [6]. This is the process for determining whether two references to real-world objects are referring to the same object or to different objects [5]. In ER, a *reference*, or *entity reference*, is a digital representation of an entity, not the entity itself. An *entity* is the actual physical real-world object; a person, place, or thing. Entity references have been referred to as entity descriptions [6], entity representations [6], objects [7], elements [8], items [8], and entity profile [9]. Entities have been referred to as an objects [6, 10], and clusters [8]. An entity reference consists of a collection of *entity attributes* that connect the reference to the real-world entity. Attributes have been referred to as features [6, 8], feature element [7, 8], and tokens [9]. With only these few terms defined, it is obvious that a standard vocabulary is needed as both reference and entity are both referred to as objects in different papers [6, 7, 10] and entity reference is defined by different terms in the same paper [8].

Other concepts that pertain to Relationship Resolution, that are not used when discussing other methods of ER, should also be identified and defined. Some of these concepts are from the study of graph theory while others are borrowed from other literature on Relationship Resolution. Graph theory defines a *graph* as a collection of vertices and a collection of edges that connect pairs of vertices. This definition holds true for graphs in Relationship Resolution. The difference is how vertices and edges are defined in this new context. Vertices are commonly referred to as nodes and edges are referred to as hyperedges[11, 12].

In Relationship Resolution, there are algorithms that act on one of two types of graphs, reference graphs and Entity-Relationship graphs. *Reference graphs* are graphs that show the connections between entity references. In a reference graph, each of the vertices represents a single entity reference and each edge represents some form of connection between a pair of references. *Entity-Relationship graphs* are graphs that show the connections between entities. In an Entity-Relationship graph, the vertices represent the entities in the dataset and the edges represent the relationships among the entities [13, 14, 15, 16]. Clustering, also referred to as partitioning [10], is the ultimate goal of Relationship Resolution; it is the means by which ER is performed in a graph environment. *Clustering* is the process of resolving which vertices in a graph are referring to the same entity. When two vertices are found to match they are grouped into the same cluster. Once all vertices are processed the resulting clusters represent the entities that are identified for the dataset. In terms of Relationship Resolution, a *cluster* is a collection of vertices that all refer to the same real-world entity. Clusters have been referred to as an entity [8], and a dataset [9]. Depending on which algorithm is being used, clustering can be performed on either an Entity-Relationship graph or a Reference graph.

In Relationship Resolution, clustering algorithms use either similarity measures, relational measures, or a combination of the two to cluster vertices or to build the edges between the vertices prior to clustering. A *similarity measure*, sometimes referred to as Featured-based similarity (FBS) methods [6] and similarity function [7], is a measure that uses approximate matching functions to measure how similar the attributes of two vertices are. If the similarity is above a predefined threshold then the vertices are added to the same cluster. A *relational measure,* sometimes referred to as distance computations [8] and structural similarity measures [17], calculates how alike two vertices are by measuring the number of edges they share to common neighbors and the distance between those neighbors. This measure focuses on the common edges in the graph instead of the attributes of the vertices themselves. A *neighbor* is a vertex that is connected to another vertex via an edge or series of edges. A *neighborhood* is a collection of vertices that are all neighbors through a predetermined number of edges or distance. There is one more term from graph theory that is well defined when referring to clustering algorithms

that can also be applied to Relationship Resolution: *Blocks* are groups of vertices that have been grouped together by a fast and cheap algorithm, typically applying some domain knowledge, into a type of candidate list. Blocks are built before the clustering algorithm is run and the aim is to reduce the number of comparisons by only applying expensive ER techniques to vertices that exist within the same block. Special types of blocks which contain overlapping vertices are referred to as canopies [8]. The process of building these blocks is called *blocking*, also referred to as match prospecting [5].

4 Classifications

With all of the different Relationship Resolution methods that have been proposed, there is no one single view that provides a universal classification system. This makes the attempt to build a classification system more difficult but not impossible. By taking into consideration different ways to view how particular methods work, three different views of classification can be defined: Measure Based, Data-type Based, and Domain Based. Each of these views can be broken down into specific classifications that are directly applicable for the specific view.

4.1 View 1: Measure Based Classification

Measure Based Classification focuses on classifying Relationship Resolution methods on the type of measure(s) that are used to identify matches between vertices in a Reference graph or Entity-Relationship graph. There are three classifications that the various methods can be divided into using the Measure Based Classification: Similarity, Relations, and Hybrid.

4.1.1 Similarity Based

As defined in the Term section of this paper, a similarity measure is a numeric measure, between 0 and 1, that is calculated by applying an approximate string matching (ASM) algorithm to the attributes of two references in the data set. This is a form of pair-wise matching since the algorithm focuses on two references at a time. If the similarity measure is above a predefined threshold then an edge is drawn between the two reference nodes in the graph. This method starts with a set of nodes and builds a graph based on the similarity between the nodes. The edges of the graph represent the similarity measure and the reference nodes that have edges with similarity measures above the threshold get merged. In [7], the author uses the attribute values of each reference to construct a context-based entity description (CED), which make up the vertices of the graph. Once the vertices are defined a similarity function is applied to the vertices and an edge is drawn between them carrying the similarity measure. Any two vertices that have a similarity above the specified threshold, denoted as Θ, are added to the same cluster. Each resulting cluster is considered a collection of references to a real-world entity.

Similarity measures that rely on ASM algorithms are performing syntactic similarity [5]. There is no one definition as to what constitutes syntactic similarity which has led to numerous ASM algorithms being defined based on how the creator views syntactic similarity. A few common algorithms are: the Levenshtein Edit Distance (LED) ASM [18], Normalized LED [19], Jaro [20], Jaro-Winkler [21], Jaccard Coefficient [22], Cosine Similarity [23], q-Gram [24], q-Gram Tetrahedral Ratio [25], and many others. Syntactic similarity, however, is not the only types of similarity measures available. There are similarity measures that rely on the semantic similarity [5], or linguistic meaning, instead of character similarity. These similarities can be based on a simple look-up table of nicknames to more complex interpretations of the attributes, discussed in [26] as latent semantic analysis. One other type of approximate matching is the use of derived match codes [5]. These algorithms are applied to attribute data to derive value that can be used to perform a deterministic match of the attributes. A few of the common derived match code functions are: Soundex [27] and the New York State Identification and Intelligence System (NYSIIS) algorithm [28].

4.1.2 Relations Based

A relational measure is a measure that only considers the connections, via edges, that are present in the graph. An algorithm is applied that traverses the graph and looks for connections between vertices that can identify them as referring to the same entity. This method of resolving entities is referred to as Association Analysis in [5]. Relation based Relationship Resolution can make discoveries that may be missed by similarity based measures if the wrong approximate string matching (ASM) algorithm is applied to the data or due to the lack of similarity in the attributes when only two references are considered. For example, through relational measures, it can be found that 'Joe Jacobs' and 'Joseph Jacob' are the same person if they both are found to have an edge that shows they are both married to 'Julie Jacobs'. In [17], the author focuses on the neighborhoods of vertices and uses a relational measure, in the form of a structural similarity measure called SCAN, which visits each of the vertices in the graph only once and detects clusters, hubs, and outliers that exist in the graph. By defining the hubs, vertices that connect multiple clusters but do not belong to any one cluster, and the outliers, vertices that connect to only one cluster but does not belong to it, in addition to the clusters, the algorithm can provide very accurate resolutions. Each cluster, hub, and outlier is considered collections of references that refer to a real-world entity. There are many different algorithms that focus on relational measures when doing clustering for ER. A few more of these relational measures are defined in [6, 10].

An example of association analysis, as shown in [5], can be seen in Fig 2. Even though none of the vertices are similar enough to merge using similarity measures, by taking into consideration the relationships between all four vertices it is easy to infer that both John Smith vertices are the same

real-world entity and both Mary Smith vertices are the same real-world entity.

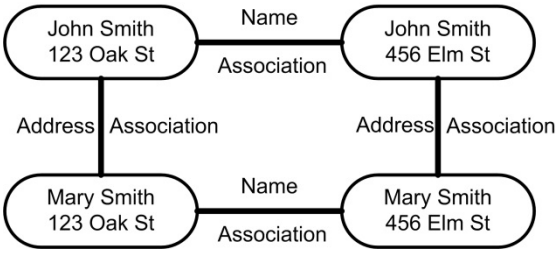


Figure 2. A simple reference graph that contains four vertices and four edges connecting the vertices.

4.1.3 Hybrid

Recently there have been some advances in ER in multi-type data. Some of the newer clustering methods use a combination of similarity measures and relational measures to get the most accurate resolutions possible. The algorithms all function differently but the basic premise is that they use the similarity measure to build an Entity-Relationship graph from the initial reference graph. Once the Entity-Relationship graph has been defined, relational measures are used to analyze the neighborhood of each node in the Entity-Relationship graph to determine if any resolutions were missed by the similarity measures. In [11], the author uses similarity measures to build an initial graph and to merge any vertices that are found to meet a predefined similarity threshold. Once the graph is built the proposed algorithm identifies the different types of data in the graph and considers the relations between the typed data to further consolidate or correct previous consolidations. Typed data can be things like an author and his/her paper; both are represented by different vertices in the same graph and each represents a different type of data. This is just one type of algorithm that incorporates both similarity and relational measures although others tend to function very similar to this one. Some other hybrid approaches are defined in [8, 9, 12].

4.2 View 2: Data-type Based Classification

Data-type Based Classification focuses on classifying Relationship Resolution methods on the type of data on which the algorithm is designed to perform Relationship Resolution. There are three classifications that the various methods can be divided into using the Data-type Based Classification: Structured, Unstructured, and Semi-Structured.

4.2.1 Structured

Structured data is the easiest form of data to analyze when performing ER. The reason for this is that structured data is well defined and all the data must be uniformly or explicitly designated. In order to be structured data each record must be clearly defined by its attribute values, and each attribute value must be labeled or positioned correctly. Data is considered structured if there is a fixed pattern or syntax that always allows the desired attribute value to be identified (repeatable i.e. last name is always in the last name field).

By definition, structured data is data that is identifiable because it has a structure or a schema. Many Relationship Resolution methods are focused on improving resolution of structured data over other types of direct matching ER solutions. In [11], the author focuses on a structured set of author information which contains the author and papers written by said author stored in a database. The author uses this attribute data as defined in a previous section to build an initial graph for their resolution. Some other methods are defined in [8, 12, 17]. Structured data comes in other forms other than databases; the data can be spreadsheets, fixed-width/delimited flat files, and XML documents. Google has also developed its own form of structured data called Bigtable [29]. It is important to understand that structured data may contain data that alone would be considered unstructured. For example, a spreadsheet may contain a "Photo" column that contains a photo of the person being referenced by the rest of the data. So long as the algorithm simply needs to know where to retrieve the photo and not how to analyze the photo for an eye color, or other physical attribute, then the data is still considered structured.

4.2.2 Unstructured

Unstructured data is one of the hardest forms of data on which to perform ER. Unstructured data has no identifiable structure which means the entity attributes are not uniformly or explicitly designated. This type of data requires specialized match functions that can parse and analyze phrases in the data for similarities. Some plagiarism software use Relationship Resolution on unstructured data to determine if parts of a submitted work were copied from other works [30]. Unstructured data also requires a Relationship Resolution method that is robust and flexible enough to incorporate the special matching algorithms required to find similarity in unstructured data. Another of these methods is defined in [10].

There are unique challenges when working with this type of data. Unstructured data consists of any data stored in an unstructured format at an atomic level. That is, in the unstructured content, there is no conceptual definition and no data type definition; in textual documents, a word is simply a word [3]. There are two overall types of unstructured data [31]. The first is *bitmap objects*; they are inherently non-language-based and include things such as images, videos, and audio files. The second type of unstructured data is *textual objects* which is based on a pre-defined language and includes things such as text documents, newspaper articles, and blog posts. Even though both of these types are understood to be data, there are very few workable models for dealing with information in bitmap objects and none efficient enough to incorporate into an ER system. Any Relationship Resolution method proposed today works with syntactic and semantic similarities in textual objects.

4.2.3 Semi-Structured

Semi-structured data incorporates traits from both structured and unstructured data. For example, in [6, 7, 9], the authors deal with the issues of linking bibliographic references. Bibliographic references are semi-structured as parts of the references can be easily identified and called out, while other parts are not easily identifiable, it greatly depends on the format it is in, MLA, APA, Chicago Style. E-mails are also semi-structured data [5] in that the heading information is all easily obtained such as the title, subject, and authors, but the content in the body are unstructured. Semi-structured data requires a Relationship Resolution method that incorporates methods for the structured and unstructured data. These types of data are typically easier to resolve as the structured parts of the documents usually provide enough information to determine initial matches which minimize the number of comparisons that have to be made by the very costly algorithms that process the unstructured data. The structured parts of the documents may also supply adequate information to allow an algorithm to more easily parse the data in the unstructured sections.

4.3 View 3: Domain Based Classification

Domain Based Classification focuses on classifying Relationship Resolution methods into the methods that require specific domain knowledge to function and the ones that can adapt to any given data set. There are two classifications that the various methods can be divided into using the Domain Based Classification: Domain Specific and Adaptive.

4.3.1 Domain Specific

A domain is the limiting factor that controls the acceptable values of the data attributes. A domain could be the "State Employees in Arkansas" which tells you that the data only pertains to employees of the state of Arkansas, or even something like "All the papers published in the ICIQ 2011 conference". In [7], the author is working in the domain of bibliographic references for major computer science journals and proceedings through the DBLP datasets. Domain knowledge is important because a domain expert knows information about the domain that can be used to optimize an algorithm that is run on the domain data. In some Relationship Resolution algorithms, domain knowledge is used to "tune" the initial reference graph by applying knowledge to build a cheap distance measure to perform blocking on the references. This domain knowledge can greatly cut down on the amount of comparisons that need to be done by more costly and generic similarity and relational measures. Some other methods are defined in [8, 9, 11, 12, 17].

4.3.2 Adaptive

Unlike domain specific algorithms, adaptive methods require a training data set to "tune" their clustering algorithms to any data set independent of any domain knowledge. This is beneficial since no domain expert is required prior to the run and the amount of human error is greatly reduced. Once the algorithm is trained it can be applied repeatedly to data of the same type. The down side to these adaptive approaches is that if the training data set does not accurately represent the data that is in the full set, the final resolutions may be skewed. Since the algorithms are adaptive, there is no way to define the data on which they are designed to work. Some of the methods are defined in [6, 10].

5 Summary

We found that although there are various concepts that are used universally throughout the literature pertaining to Relationship Resolution, there was no specific vocabulary defined to refer to these concepts. This is common in new fields of study and allowed multiple authors to refer to the same concept using different terms. We attempted to clarify some of the common concepts but believe that there is a need and opportunity for someone to define a base vocabulary consisting of concepts applied in ER and Relationship Resolution. That is beyond the scope of this paper.

Also, even though there are many different views as to what Relationship Resolution is, we found that the works can be classified if considered from three separate viewpoints: Measure Based, Data-type Based, and Domain Based. Each view was further broken down into classifications. Although there was no way to define an overall trend in the algorithms, it was possible to see how the various algorithms were distributed when broken down among the views.

By decomposing the Measure Based view into similarity measures, relational measures, and hybrid approaches, we saw that the majority of Relationship Resolution algorithms that are being proposed use a hybrid approach which uses a combination of the similarity measure and the relational measures to improve resolution decisions. By categorizing the Data-type Based view into structured, unstructured, and semi-structured data, we see that most algorithms depend on structured data to perform clustering and resolution. There is, however, a growing interest in performing ER on bibliographic data which falls into the realm of semi-structured data; these are a more narrowed focus set of algorithms. Lastly, we broke the Domain Based view into domain specific and adaptive classification. We saw that an overwhelming majority of algorithms for Relationship Resolution are Domain specific and still require a domain-expert to tune the algorithms for the specific data sets on which they run.

Through our observations, we find that there is a gap in the available algorithms for performing Relationship Resolution on unstructured textual objects and bitmap objects. There are more available algorithms for unstructured textual objects but there is little work available on performing Relational Resolution on bitmap objects.

6 Future Work

As more research is done into Relationship Resolution, it will be possible to further categorize the various methods into their specific domains. This is important since different algorithms are inherently better suited for specific types of information. By further classifying the algorithms into their specific domains, it may allow for further research between similar algorithms so that the functionality can be incorporated and the algorithms can be improved through the findings of similar works.

Acknowledgment

The research described in this paper has been supported in part by research grants from the Arkansas Department of Education and the National Institutes of Health Clinical and Translational Science Awards program.

References

[1] Welcome to ORCID., *Welcome to ORCID.*, [online] 2012, http://about.orcid.org/ (Accessed: 10 Mar. 2012).

[2] G. Navarro. "A guided tour to approximate string matching," *ACM Computing Surveys*, Vol. 33, Issue 1, 31-88, 2001.

[3] F. Naumann, *Quality-Driven Query Answering for Integrated Information Systems*, Berlin: Springer-Verlag, 2002.

[4] F. Naumann and M. Herschel, "An introduction to duplicate detection," *Synthesis Lectures on Data Management*, Vol. 2, no. 1, pp. 1-87, 2010.

[5] J.R. Talburt. *Entity Resolution and Information Quality*, Burlington, MA: Morgan Kaufmann, 2010.

[6] Z. Chen, D.V. Kalashnikov, and S. Mehrotra, "Adaptive graphical approach to entity resolution," in *Proc. 7th ACM/IEEE-CS joint conf. on Digital libraries (JCDL '07)*, New York, NY, USA, ACM, 2007, pp. 204-213.

[7] L. Li, J. Li, H. Wang, and H. Gao, "Context-based entity description rule for entity resolution," in *Proc. 20th ACM Int. Conf. on Information and Knowledge Management (CIKM '11)*, Bettina Berendt, Arjen de Vries, Wenfei Fan, Craig Macdonald, Iadh Ounis, and Ian Ruthven (Eds.), New York, NY, USA, ACM, 2011, pp. 1725-1730

[8] A. McCallum, K. Nigam, and L.H. Ungar, "Efficient clustering of high-dimensional data sets with application to reference matching," in *Proc. 6th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD '00)*, New York, NY, USA, ACM, 2000, pp. 169-178.

[9] G. Papadakis, E. Ioannou, C. Niederée, and P. Fankhauser, "Efficient entity resolution for large heterogeneous information spaces," in *Proc. 4th ACM Int. Conf. on Web Search and Data Mining (WSDM '11),* New York, NY, USA, ACM, 2011, pp. 535-544.

[10] B. Long, Z.M. Zhang, and P.S. Yu, "A probabilistic framework for relational clustering," in *Proc. 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD '07)*, New York, NY, USA, ACM, 2007, pp. 470-479.

[11] I. Bhattacharya and L. Getoor, "Relational clustering for multi-type entity resolution," in *Proc. 4th Int. Workshop on Multi-relational Mining (MRDM '05)*, New York, NY, USA, ACM, 2005, pp. 3-12.

[12] I. Bhattacharya and L. Getoor, "Collective entity resolution in relational data," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, Article 5, March 2007.

[13] D. V. Kalashnikov, S. Mehrotra, and Z. Chen. "Exploiting relationships for domain-independent data cleaning," in *Proc. SIAM Data Mining (SDM) Conf.*, 2005, pp. 1-48.

[14] B. Malin, "Unsupervised name disambiguation via social network similarity," in *Proc. Workshop on Link Analysis, Counterterrorism, and Security*, 2005 SIAM International Conference on Data Mining, Newport Beach, CA, 2005, pp. 93-102.

[15] E. Minkov, W.W. Cohen, and A.Y. Ng, "Contextual search and name disambiguation in email using graphs," in *Proc. 29th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, New York, NY, USA, ACM, August 2006, pp. 27-34.

[16] B.-W. On, E. Elmacioglu, D. Lee, J. Kang, and J. Pei, "Improving grouped-entity resolution using quasi-cliques," in *Proc. 6th Int. Conf. on Data Mining (ICDM '06)*, Washington, DC, USA, IEEE Computer Society, 2006, pp.1008-1015.

[17] X. Xu, N. Yuruk, Z. Feng, and T.A.J. Schweiger, "SCAN: a structural clustering algorithm for networks," in *Proc. 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD '07)*, ACM, New York, NY, USA, ACM, 2007, pp. 824-833.

[18] V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, vol.10, Issue 8, pp. 707-710, 1966.

[19] P. Christen. "A comparison of personal name matching: techniques and practical issues," in *Proc. 6th IEEE Int. Conf. on Data Mining-Workshops (ICDMW'06)*, Hong Kong, pp. 290-294, 2006.

[20] M. Jaro, "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida," *Journal of the American Statistical Association*, Vol.84, Issue 406, pp. 414-420, 1989.

[21] W. Winkler. "The State of Record Linkage and Current Research Problems," Statistical Research Division, U.S. Census Bureau, 1999.

[22] A. Monge and C. Elkan, "The field matching problem: algorithms and applications," in *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining*, 1996, pp. 267-270.

[23] P.-N. Tan, M. Stenbach, and V. Kumar, *Introduction to Data Mining*, Boston, MA: Addison-Wesley Longman Publishing Co. Inc., 2005.

[24] E. Sutinen and J. Tarhio, "On using q-gram locations in approximate string matching," in *Proc. 3rd Ann. European Symp. on Algorthims (ESA '95)*, London, UK, Springer-Verlag, 1995, pp. 327-340.

[25] G. Holland and J. Talburt, "q-Gram tetrahedral ratio (qTR) for approximate string matching," in *Proc. 2010 Ann. Acxiom Laboratory for Applied Research Conf. (ALAR-10)*, 2010

[26] R. Deaton, T. Doan, and T. Schweiger, "Semantic data matching: Principles and performance." in *Data Engineering: Mining, Information and Intelligence*, Y. Chan, J. Talburt & T. Talley (Eds.), Vol. 132, Springer, pp. 77-90, 2010.

[27] M. Odell and R. Russell, U.S. patent number 1,261,167, Washington, D.C: U.S. Patent Office, April 2, 1918.

[28] R.L. Taft, "Name Search Techniques, New York State Identification and Intelligence System," *Special Report No. 1*, Albany, New York, 1970.

[29] F. Chang, J. Dean, S. Ghemawat, W.C. Hsieh, D.A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R.E. Gruber, "Bigtable: A distributed storage system for structured data." In *Proc. 7th USENIX Symp. on Operating Systems Design and Implementation - Volume 7 (OSDI '06)*, vol. 7, Berkeley, CA, USA, USENIX Association, 2006, pp. 15-15.

[30] C. Liu, C. Chen, J. Han, and P.S. Yu, 2006. "GPLAG: detection of software plagiarism by program dependence graph analysis," in *Proc. 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD '06)*, New York, NY, USA, ACM, 2006, pp. 872-881.

[31] M.J. Crowsey, A.R. Ramstad, D.H. Gutierrez, G.W. Paladino, K.P. White, "An evaluation of unstructured text mining software," *Systems and Information Engineering Design Symposium, 2007. SIEDS 2007. IEEE* , pp.1-6, 27 April 2007.

Developing and Refining Matching Rules for Entity Resolution

Huzaifa Syed, John Talburt, Fan Liu, Daniel Pullen and Ningning Wu
Information Science Department, University of Arkansas at Little Rock
UALR EIT 550, 2801 S. University Ave., Little Rock, AR 72204

Abstract - *This paper describes a process for developing and refining matching rules for Entity Resolution. The process is an iterative scheme that has two phases. The first phase tries to identify the primary entity identity attributes and baseline matching rules. The second phase consists of identifying and incrementally augmenting the baseline rules with supporting attributes in a way that incrementally reduces the number of false positive and false negative resolutions and ultimately brings the resolution results closer to the truth. The process design was implemented in the open source entity resolution system named OYSTER.*

Keywords: Entity Resolution Rules, Developing Rules in Entity Resolution, Refining rules in Entity Resolution, OYSTER, Support Rules

1 Background

Entity Resolution (ER) is the process of determining whether the two references to real world objects in an information system are referring to the same object or two different objects [1]. Entities are described in terms of their characteristics, called attributes. A reference is a collection of attribute values for a particular entity.

Most ER processes still follow the Fellegi-Sunter Model for record linking [2] in which pairs of records are judged to be "link", "non-link" pairs, or perhaps "possible link" pairs depending upon which attribute values agree or disagree. For example, the pattern that two student enrollment records agree on first name values, agree on last name values, agree on date-of-birth values, but disagree on school identifier values, might be designated a link rule. This is the decision that the records are for the same student. Matching rules that give a yes or no (true or false) decision are called Boolean match rules [3].

As an example, consider the student records shown in Table 1 where there are four attributes. "RefID" represents the unique identifier for each record. "FirstName" represents the student's given name. "LastName" represents the student's surname. Lastly, "DOB" represents the student's date-of-birth in the format Month/Day/Year.

Suppose that a Boolean matching rule using these attributes is defined as follows:

IF
    FirstName values are an Exact Match (identical)
AND
    DOB values are an Exact Match
THEN
    Link the records

RefID	First Name	Last Name	DOB
1	John	Doe	3/19/1980
2	John	Doeson	3/19/1980
3	Daniel	Gilburt	3/19/1990
4	Daniel	Gilbert	3/19/1990
5	John	Doe	6/29/1960
6	Joseph	Doe	5/19/1981

Table 1: Records to resolve

An ER process using this matching rule will produce the results shown in Table 2. Here Records 1 and 2 are linked as are Records 4 and 5. The linking relationship is shown by the value of a new attribute added to each record called the "LinkID". The values of the LinkID represent the decisions of the ER process in that two records will have the same link ID if and only if the ER process has decided that the references are for the same student according to the logic of its matching rules.

RefID	First Name	Last Name	DOB	Link ID
1	John	Doe	3/19/1980	1
2	John	Doeson	3/19/1980	1
3	Daniel	Gilburt	3/19/1990	2
4	Daniel	Gilbert	3/19/1990	2
5	John	Doe	6/29/1960	3
6	Joseph	Doe	5/19/1981	4

Table 2: Resolved Records

For the process of Entity Resolution, OYSTER (Open sYSTem Entity Resolution) will be utilized. OYSTER is an ER system developed by the ERIQ Research Center at the University of Arkansas at Little Rock (ualr.edu/eriq). OYSTER provides access to a variety of entity resolution algorithms that enables users to uncover duplicate and redundant entity references [5]. OYSTER was run in Identity Capture Mode. Identity capture is another form of ER in which the system builds (learns) a set of identities from the references it processes [7]. Identity capture is essentially a "smart" version of merge-purge or record linking in which the knowledge gained in resolving the references is retained in the form of EIS [6].

2 Fellegi-Sunter Rule Structure

The Boolean matching rules discussed in this paper are based on the Fellegi-Sunter model. In this model there are four components.

Attribute: Entities are described in terms of their characteristics, called attributes [1]. The values of these attributes provide information about a specific entity. Identity attributes are those that when taken together distinguish one entity from another. Matching rules in ER operate on these attributes.

Comparator: An algorithm for determining the degree of similarity between two attribute values. Comparators are also called similarity functions [8].

Term: A logical proposition formed by a comparator, two values of an attribute, and an explicit or implied threshold value. The term is true if similarity between the two values as determined by the comparator rises to the level set by the threshold value.

Rule: A Rule is a logical proposition formed by a single Term or by the logical conjunction (AND operation) of two or more Terms [3].

Rule Set: A Rule set is a logical proposition formed by a single Rule or by the logical disjunction (OR operation) of two or more Rules.

In ER systems implementing Fellegi-Sunter rules, the decision whether two records linked or not linked is determined by whether the attribute values in two records make the rule set proposition true or false, respectively. By the definitions above, it means that the attribute values in the two records must make all of the terms true in at least one rule of the rule set. The decision not to link only happens in the case that none of the rules is satisfied.

3 Problem

There are many challenges associated with developing the rules for the ER process. The result is directly dependent upon the rules. Throughout this process, either negative or positive will be utilized to identify a true or false outcome as being correct or incorrect. The goal is to create a set of rules which maximizes the number of true positives and true negatives while minimizing false positives and false negatives. The determination of correctness is based on verification against the real-world objects. A true positive is an ER result that correctly linked two records together [1]. A false positive is an ER result that incorrectly linked two records together [2]. A true negative is a pair of references that reference different objects and were not linked by the ER process. A false negative is an ER result where a pair of references that refer to the same object were not linked by the process. Table 3 shows a summary of these measures and their relationships in a confusion matrix. P (positive) and F (false) are the total of all correct positives and negatives respectively. P' and N' are the total of all positives and negatives as predicted by the ER process.

		Correct Decision		
		Should Link	Should Not Link	**Total**
ER Process	Linked	TP	FP	P'
	Not Linked	FN	TN	N'
	Total	P	N	

Table 3 Confusion Matrix

In pursuit of producing more true positives, more false positives and false negatives are typically produced. For example, consider S, a set of records, displayed in Table 4 below. ER needs to be performed on S. To accomplish this, make a rule that will match exact on SSN, exact on Last Name and exact on DOB. To satisfy an exact match condition, the comparison between the two records may not deviate by case or contents. When these rules are applied, it will make true positives for Joey and Joseph but will also make false positives for George, Stanley and a false negative for George Doe. Comparing the result with the true information, it is possible to deduce the correct linking.

SSN	First Name	Middle Name	Last Name	DOB
779-889-9999	George	S	Doe	8/9/1975
779-889-9999	Stanley	Lisa	Doe	8/9/1975
963-456-8932	Joseph		Gilbert	5/29/1965
963-456-8392	Joey		Gilbert	5/29/1965
775-899-9999	George	Steve	Doe	8/9/1975

Table 4: Record Set S

After the ER process, the result summary will look something like the Venn diagram in Figure 1. The square field of Figure 1 represents SxS, all possible pairs of records. The vertical oval represents the pairs of records that the ER rule has linked together while the horizontal oval labeled ExE represents the pairs of records that are actually equivalent, i.e. refer to the same student. The areas labeled FP are where the rule linked non-equivalent records (false positives) and the areas labeled FN are equivalent pairs that were not linked (false negatives).

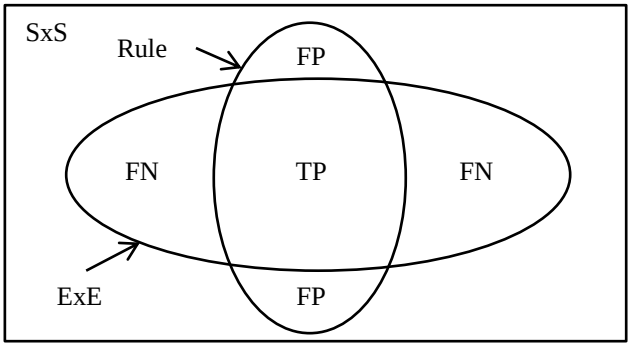


Figure 1: Single Rule

However, it is rare that a single identity rule will be effective. In most cases, a single rule will be too coarse,

producing too many false positives and false negatives. The ratio of the area TP to the total area of ExE is the "recall" of the rule, and the ratio of the area of TP to the total area of the rule oval is the "precision" of the rule. The goal of rule development is to craft a set of rules where each rule has high precision, and the total recall of all rules is high as shown in Figure 2.

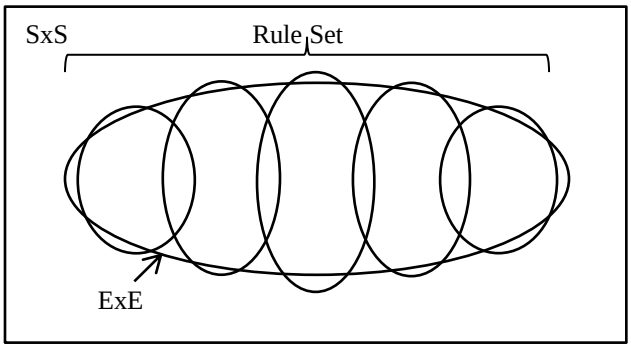


Figure 2: Rule Refinement

4 Solution

To improve the results, it is necessary to analyze where the tested rule set differs from the true identities. For large data sets this is only practical for some sample of the data. The biggest problem in the analysis process is finding the false negative errors of the rule. Inspecting the records that were clustered (linked) together by a rule may help discover false positive errors, but by definition, the records comprising false negative errors will not be in the cluster. Those records could be anywhere in the data set.

In addressing the false negative analysis, it is very helpful to have some type of benchmark to guide the analysis. An ER benchmark typically has one of two forms. It could either be a subset of the records for which the correct linkage is known (a truth set), or it could be prior linking of the entire set of records that experience has shown is close to the truth.

Based on the analysis, the rule or rules are refined, and the process is repeated until the rules meet the acceptable threshold of error for the application. Typically the process begins with a single rule as shown in Figure 1. However, during the analysis process, additional supporting rules that will further aid in approaching the truth may be defined. As shown in Figure 2.

Another important consideration in the rule development process is data quality. The first step in the process is to understand what attributes are in the data and the condition of the values for each attribute. This is best done by data profiling. The profiling of each attribute provides important information about whether the attribute has a complete set of values (null or blank values), the uniqueness of the values, and how closely the values follow any data standards or validity rules that are in place.

It is often the case that the data requires cleaning and standardization before the rule development process can begin. Typical processes include consistent letter casing (usually all uppercase), removal of punctuation and special characters, standardization of names and abbreviations, and other consistent formatting.

The data in the study presented here was a large set of student data spanning multiple years. The same student would not only have records in different years, student changing schools within the same school district would have multiple records within the same school year. The data also presented many data quality problems due to data entry errors, differences in formatting from year-to-year and school-to-school, and situational changes, such as changes in name and school identifiers.

Phase One:

Defining rules is an iterative process. The goal is to keep refining the identity rules until the false negative and false positive errors that they produce fall with an acceptable limit. In order to refine the rules, the similarity between the truth and the ER result must be measured. At the same time, the complete truth is unknown. Otherwise, there would be no need for performing the ER process. Fortunately, in most cases there is a benchmark that is known, or assumed to be, reasonably close to the truth. This might be links that have been created by past processes, either manual or automated.

For this discussion, the benchmark link value is called PriorLink. This proprietary identifier historically was considered to be close to the truth. In this study, PriorLink was used as the benchmark and the identity rules were measured against it. The measurement of the overall similarity between the linking given by PriorLink and the linking of any given set of identity rules was determined by the Talburt Wang Similarity Index (TWi). The TWi uses the number of overlaps between two set of partitions (in this case record clusters) and calculates an index which represents the similarity between the two partitions [4]. The TWi of similarity between two partitions A and B is defined as:

$$TWi = \frac{sqrt(|A| \cdot |B|)}{|V|}$$

This is where A and B are two partitions of a set S, and V is the set of overlaps between A and B.

Table 5 below is a test report. This report is primarily a comparison between the rules in Figure 7 against the PriorLink benchmark.

The rule defined for this run in OYSTER is displayed in Figure 3 below.

```
<IdentityRules>
  <Rule Ident="FLD”>
        <Term Item="StudentFirstName" MatchResult="Exact"/>
        <Term Item="StudentLastName" MatchResult="Exact"/>
        <Term Item="StudentDOB" MatchResult="Exact"/>
  </Rule>
</IdentityRules>
```

Figure 3: Initial Rule set 1

	A	B	C	D	E	F
1		OYSTER Record Linking Test Report				
2		ADE Project				
3	Test Date:	Fri Feb 24 14:04:21 CST 2012				
4	Test Objective:	Evaluate performance of OYSTER rules against PriorLink Identifier				
5	OYSTER Rule Set:					
6	Test DataSet:					
7		PriorLink Versus OYSTER				
8						
9	Statistic	PriorLink	OYSTER			
10	Total Records	98276	98276			
11	TW Index	1	0.999054			
12	Clusters Count	97098	97126			
13	Overlaps	97098	97204			
14	Average Cluster Size	1.012132073	1.01184			
15	Cluster size Distribution					
16	Cluster Size	Count	Count			
17	1	95931	95990			
18	2	1156	1122			
19	3	11	14			
20	4	0	0			
21	5	0	0			
22	6	0	0			
23	7	0	0			
24	8	0	0			
25	9	0	0			
26	10	0	0			
27	10+	0	0			
28	MAX	3	3			

Table 5: Test Report

From the above Test Report, the similarity between the rule set and PriorLink identifier is 99.90%. To further improve the results, they must be evaluated and analyzed. Compare the clusters of PriorLink with those of OYSTER, then identify and analyze the cases where PriorLink's clustering is splitting that of the OYSTER results and where OYSTER's clustering is splitting the clustering of PriorLink. By looking at the cases where PriorLink is bringing two records together and the rules are not, the cause of the errors produced by the OYSTER rules can be identified. This allows for the rule set to be improved for the next run. By looking at the cases where OYSTER is splitting PriorLink, the potential false positives by the rules and false negatives by the benchmark (PriorLink) can be identified. Hence, the rules have the potential to be made better than the benchmark. To facilitate this process, a tool was developed to automatically compare the result of two OYSTER runs. Its result includes a report like that in Table 5. It also includes two reports that show the cases where PriorLink splits the OYSTER rules and the OYSTER rules split PriorLink as identified above. One of these reports can be seen in Table 6. The full list of available attributes has been limited to those in the current discussion. This split comparison allows for the conclusions to lead into phase two.

Phase Two:
By looking at the split of PriorLink vs. OYSTER (i.e. cases where the rule was not able to bring two records together and PriorLink was able to cluster them together), the shortcomings of the rules were identified. To overcome the shortcomings, a rule in support of the basic rule was defined.

First Name (Exact), Last Name(Exact), DOB(Exact)

	Oyster ID	PriorLink	First Name	Last Name	DOB
Source1.16152	0L2O22XH4Z3B200Y	1234567890	BOB	SMITH	1/21/1995
Source1.12144	0L2O22XH4Z3B200Y	1234567891	BOB	SMITH	1/21/1995
Source1.324646	0LBRJZF32F9TRWB4	1234567892	JOE	BLOW	2/27/1995
Source1.489238	0LBRJZF32F9TRWB4	1234567892	JOE	BLOW	2/27/1995
Source1.497887	0LBRJZF32F9TRWB4	1234567893	JOE	BLOW	2/27/1995
Source1.410463	0QISMBJ2OP9DGEOQ	1234567894	JIMMY	JOHN	10/1/2001
Source1.410848	0QISMBJ2OP9DGEOQ	1234567895	JIMMY	JOHN	10/1/2001

Triand

	Oyster ID	PriorLink	First Name	Last Name	DOB
Source1.16152	UEH1S376J1EVTI6F	1234567890	BOB	SMITH	1/21/1995
Source1.12144	NAIRD48G2PTPR463	1234567891	BOB	SMITH	1/21/1995
Source1.324646	15KJBZ6AVFRFGOTR	1234567892	JOE	BLOW	2/27/1995
Source1.489238	15KJBZ6AVFRFGOTR	1234567892	JOE	BLOW	2/27/1995
Source1.497887	83NS8BLJF8QWT7NE	1234567893	JOE	BLOW	2/27/1995
Source1.410463	PJH4WQ1NKRY1V0PW	1234567894	JIMMY	JOHN	10/1/2001
Source1.410848	6TDHPPT0HNQDZROA	1234567895	JIMMY	JOHN	10/1/2001

Table 6: Clustering Comparison

For example, a case where the current rule was not bringing two records together was identified. One cause was typographical errors in the Date of Birth of several entities. By implementing the supporting rule it is possible to eliminate many false negatives. In addition to typographical errors in the data of birth field, inconstancies in student names were found. This inconsistencies were typically the use of a nickname in one record while using the full given first name in another record. It is necessary to also look at the cases where OYSTER was splitting PriorLink (i.e. cases where the rule was able to bring two records together and PriorLink was not able to). This identified the cases where PriorLink was producing a false negative. This identification shows the OYSTER rules were outperforming the PriorLink benchmark on at least this sample of records. This also gives the opportunity to look at the false positives by the rule which again provides the opportunity for improvement of the rules.

Table 7 below is one of the latest test reports. In this test report, the rules in Figure 4 are compared against the benchmark, PriorLink. The TWi similarity index is 99.96%. It shows an increase from the TWi of Table 5. It can be seen that many support rules are added to the base rule. The support rules were added to overcome the limitations of the base rule. The refined set of rules as defined in OYSTER are shown in Figure 4 are produce fewer false positives and negatives.

	A	B	C	D	E
1		OYSTER Record Linking Test Report			
2		ADE Project			
3	Test Date:	Fri Mar 09 16:28:51 CST 2012			
4	Test Objective:	Evaluate performance of OYSTER rules against PriorLink			
5	OYSTER Rule Set:				
6	Test DataSet:				
7		PriorLink Versus OYSTER			
8					
9	Statistic	PriorLink	OYSTER		
10	Total Records	98276	98276		
11	TW Index	1	0.999624		
12	Clusters Count	97098	97155		
13	Overlaps	97098	97163		
14	Average Cluster Size	1.012132073	1.011538		
15	Cluster size Distribution				
16	Cluster Size	Count	Count		
17	1	95931	96045		
18	2	1156	1099		
19	3	11	11		
20	4	0	0		
21	5	0	0		
22	6	0	0		
23	7	0	0		
24	8	0	0		
25	9	0	0		
26	10	0	0		
27	10+	0	0		
28	MAX	3	3		

Table 7: Test Report

```
<IdentityRules>
   <Rule Ident="FLD">
         <Term Item="StudentFirstName" MatchResult="Exact"/>
         <Term Item="StudentLastName" MatchResult="Exact"/>
         <Term Item="StudentDOB" MatchResult="Exact"/>
   </Rule>
   <Rule Ident="FLDa">
         <Term Item="StudentFirstName" MatchResult="QTR(0.25)"/>
         <Term Item="StudentMiddleName" MatchResult="Initial"/>
         <Term Item="StudentLastName" MatchResult="Exact"/>
         <Term Item="StudentDOB" MatchResult="Exact"/>
         <Term Item="ParentFName" MatchResult="Exact"/>
   </Rule>
   <Rule Ident="FLDb">
         <Term Item="StudentFirstName" MatchResult="Exact"/>
         <Term Item="StudentLastName" MatchResult="QTR(0.25)"/>
         <Term Item="StudentDOB" MatchResult="Exact"/>
         <Term Item="ParentFName" MatchResult="Exact"/>
   </Rule>
   <Rule Ident="FS">
         <Term Item="StudentFirstName" MatchResult="Exact"/>
         <Term Item="SSN" MatchResult="Exact"/>
   </Rule>
   <Rule Ident="LS">
         <Term Item="StudentLastName" MatchResult="Exact"/>
         <Term Item="SSN" MatchResult="Exact"/>
   </Rule>
</IdentityRules>
```

Figure 4: Rule Set

5 Conclusion

The ER rule creation process is an iterative process that may take many repetitions of the steps described above in Phase 2. As in any improvement process, it is critical to have periodic measurements as a way to gauge progress. In this case there are two types of metrics, one for overall variance and another for specific differences. For this case study, the overall variance metric was the TWi selected because of its simple calculation and easy to interpret numerical rating. While understanding the overall variance is important, the TWi is not enough to evaluate an ER result. Although it gives a succinct measure of overall variance between the results, it doesn't give a direct indication as to whether the differences are more correct or less correct.

The second measure is the actual error count for some sample of linking differences between those given by the ER rules set and those given by a previously established benchmark. In performing this analysis it is necessary to inspect some sample of the actual differences, i.e. the cluster splits. Such an inspection serves two purposes. One is to evaluate the performance of the rule set against the established benchmark. A second is that the inspection of data can often provide insight into how certain rules might be modified or how new rules might be added to further improve the accuracy of the process. The manual inspection and verification is necessary because the benchmark is not itself 100% correct. The inspection process may also identify cases where the benchmark itself has produced false positives or false negatives.

6 Future Work

The refinement of ER by decomposition into supporting rules show promise for reducing false negative rates. At the same time, they can also introduce false positives. One future research direction is the development algorithms or other techniques that could assist users in identifying and evaluating candidates for supporting rules.

7 Acknowledgement

Funding for the research in this paper was provided through research grants by the Arkansas Department of Education.

8 References

[1] John Talburt. "Entity Resolution and Information Quality". Morgan Kaufmann, 2010.

[2] Iven Fellegi and Alan Sunter. "A Theory for Record Linkage"; Journal of the American Statistical Association, Vol. 64 No. 328, 1183-1210, 1969

[3] Steven Whang and Hector Garcia-Molina. "Entity Resolution with Evolving Rules"; Proceedings of the VLDB Endowment, Vol. 3 Issue 1-2, 1326-1337, September 2010

[4] John Talburt, Emily Kuo, Richard Wang and Kimberly Hess. " An Algebraic Approach to Quality Metrics For Customer Recognition Systems"; Proceedings of the Ninth International Conference on Information Quality(ICIQ-04)

[5] Zhou, Y. and Talburt, J. (2011). Entity Identity Information Management (EIIM). International Conference on Information Quality (ICIQ-11), Adelaide, Australia, November 18-20, 2011, pp. 327-341

[6] Zhou, Y., Talburt, J., Su, Y. and Yin, Y. (2010). OYSTER: A Tool for Entity Resolution in Health Information Exchange. The 5th International Conference on the Cooperation and Promotion of Information Resources in Science and Technology (COINFO'10), Beijing, China, November 27-29, 2010, pp. 356-362

[7] Zhou, Y. and Talburt, J. (2011). The Role of Asserted Resolution in Entity Identity Management. The 2011 International Conference on Information and Knowledge Engineering (IKE'11), Las Vegas, Nevada, July 18-20, 2011, pp.291-296

[8] Naumann, F. and Herschel, M. (2010). An Introduction to Duplicate Detection. Synthesis Lectures on Data Management 2010 2:1, 1-87

Information Quality Assessment and Improvement of Student Information in the University Environment

Melody Penning and John R. Talburt
Information Science Department, University of Arkansas at Little Rock

Abstract - ***This paper describes a project to assess and improve the quality of student information at a state university which combines process control and improvement with error detection and correction. A major emphasis of the project was an entity resolution process designed to consolidate redundant student records. Establishing a method to monitor data quality over time in order to support a quality process that responds to changes in the environment was a secondary focus. The project resulted in improved error detection and the development of a data quality metadata warehouse as well as a process to automate, save, and display successive data quality assessments.***

Keywords –Entity Resolution, Quality Assessment, Metadata Warehouse

I. INTRODUCTION

The main objective of this project, which tied into an existing data quality effort, was to compare the data quality of the university student dataset before and after the initiation of a data quality assessment process. It tested the effectiveness of data quality management techniques in general and record linking specifically to see the effect on the quality of the student data in the university database after the intervention. Several years of linkage measurements were available and provided the before view of the data quality. The introduction of OYSTER, entity resolution software, and a data quality metadata warehouse (DQMDW) are the interventions. The data quality was measured by counting the identified errors since this is comparable to the data maintained from prior years. The addition of a DQMDW contributes to improved quality governance because it provides a method to maintain a record of successive quality measurements for comparison and will support a process that includes measuring, saving and displaying the information quality results over time. The Total Data Quality Management (TDQM) methodology recalling the Deming cycle of a continuous circular quality process suggests that quality management makes optimal progress using a feedback loop. To that end, the final product of this project is a process anchored in the cyclic approach to quality advocated by Deming, Lee, and Wang [13][4][1] the results of which are made available to both business and IT users through a visualization dashboard [4].

II. BACKGROUND: EXISTING QUALITY STATUS

By assisting the university Data Integrity team with quality issues in the student data subsection of the Banner database the new interventions could be woven into the existing project structure. Banner is the higher-education software system used at the university. The Banner database has a combined table and view count exceeding a thousand and is incredibly complex. It is the creation of an evolutionary process spanning decades and corporations so it cannot be tackled as a whole. The student subsection is, however, a key component and can provide a template for the remaining functional database areas. General data quality was to be addressed and specifically, there was a need to correct entity resolution issues. The data comes into the database by many processes, including: university employees, students self-reporting, and batch uploads from institutions like ACT. This creates a situation where entity identity is difficult to monitor and maintain. Fortunately, the Banner software system contains tools to address data quality issues and there are also data quality scripts written specifically for use with the Banner system.

1. Banner Tools

The Banner product features entity resolution components that can be used to look up existing records and present possible matches with incoming data. The entity resolution component of the Banner system called Common Matching checks for existing matches on insert and update database interactions. The matching rules it uses are customizable and can use any number of the fields made available by the common matching component.

Scripts are also used and possible duplicates, when located, are added to an email that is sent to members of the data integrity team so that the entry can be reviewed. The Banner tools provide the raw materials to begin to explore the quality of the data but once a possible error is detected the data quality work actually begins.

2. The Data Integrity Project

The Data Integrity Project was introduced at the University of Arkansas at Little Rock in October of 2009 in order to locate and address data quality problems, primarily problems stemming from entity resolution issues. These were referred to as MPOPs by the team members. MPOP stands for *M*ultiple *P*IDM *O*ne *P*erson; a PIDM is a key number used within the Banner system to maintain entity identity. Team members were recruited from diverse campus areas; info tech services, records and registration, admissions and financial aid, financial services, recruitment and retention, and international areas. Standards were developed and, in a fashion typical in other data quality projects, the Data Integrity team focused on two main areas; prevention and clean-up [2][3]. They had an important ingredient for a successful data quality project, buy-in from upper administration [4]. This team put a process for data quality

monitoring and correction in place and has made substantial progress on the data quality front despite the enormous size of the undertaking. The project directives are: Identify MPOP, Merge MPOP, Prevent MPOP, and Tracking Project Progress.

III. Extending the Existing Data Quality Project

The extent of the Banner tools matching capability, while very useful, does not match the capability of tools built specifically for this purpose. The Data Integrity team was in need of more advanced data quality tools to address the MPOP problem, an entity resolution issue. To supplement the existing entity resolution capabilities of the university open source entity resolution software, OYSTER, was used to detect hidden MPOP errors. In order to maintain the structure of the prior measurements taken during the data integrity project, the data quality assessment and improvement addition has been conducted from the data perspective. The 'data perspective' refers to the IQ problems that exist objectively in the database in contrast to those that are only apparent from the user perspective [3]. This was necessary in order to compare the measurements from before to after the project intervention. This was also a means to limit the size of the already large and complex project in order to deliver tangible results more quickly.

1. Unresolved Data Quality Issue: MPOP

A. Entity Resolution: Entity resolution is the process of identifying which data is referring to which entity in the real world. As in the case of two database entries with the same first and last name combination, these could be referring to one person or two different people. What we're interested in is how to determine to whom the data is referring. Over the last 50 years successively more abstract models of entity resolution have been developed [11] and the application has been built into OYSTER

B. Entity Resolution applied to the MPOP problem: The problem of different id numbers associated with one person is an example of the entity resolution problem discussed above. When database entries aren't unique, two people can be merged into one or; conversely, if the entries are unique then one person can be split into two. The latter scenario is the MPOP situation. Entries are made and the misspelling of a name, or some other attribute variation, produces a difference that prevents one person's data from linking together as one identity. Correctly grouping student information together with a student identity number is the challenge.

Keeping track of all of the variations of an identity helps to link incoming information together. To say it another way, the more you know about someone the more likely you are to be able to distinguish them from other similar individuals. So if two records in a database are Lisa Johnson and Lisa Lewis, these may seem like two different people unless your Lisa's friend and are aware that she recently married Joe Lewis. This turns out to be a valuable tool for doing entity resolution called transitive closure. OYSTER, the entity resolution software discussed in the next section, provides a means to put this tool to work on the MPOPs in the university database.

C. OYSTER: The open system entity resolution system developed at the Center for Advanced Research in Entity Resolution and Information Quality (ERIQ) features a broad set of matching capabilities [11].

The most common method of record linkage is *direct matching*; this is the method that Banner uses in the Common Matching component. Comparing the attributes to one another for an exact or fuzzy match are types of direct matching.

Transitive equivalence takes direct matching a step further by linking records based on a chain of directly matched record pairs, i.e. if A matches B, and B matches C, then A and C are equivalent even if A and C do not match.

Asserted equivalence is another matching tool available in OYSTER. It allows the use of external knowledge to be included in the ER process by declaring that two records are equivalent even if they are not linked by direct matching or transitive equivalence.

OYSTER has five modes to choose from; record linking, identity capture, identity resolution, and identity update [15]. The modes determine whether data will be saved for matching with new incoming data, matched against an existing dataset and captured in the event of a match, saved and merged into an existing dataset, or simply compared for matches but not saved for future runs. A script was created to automate the execution of OYSTER and save the results to the DQMDW.

2. Unresolved Data Quality Issue: Generic

In addition to the entity resolution, MPOP, quality issue other categories of quality errors such as accuracy, consistency, and completeness were captured as well [5][8][4]. These IQ problems can all be identified using logic applied through SQL. Attribute domain constraints like optionality, format, and valid value were used along with relational integrity rules to search for errors [8]. To facilitate automation [4] these queries were grouped together in a script; however, the process to find these types of errors is the same whether they are grouped together in one script or as individual SQL files. All of the IQ errors, both generic and MPOP, were added to the DQMDQ for verification and for comparison over time.

3. DQMDW

Retaining and managing quality metadata has been addressed by many authors in an effort to improve both quality and processes [6][8][2][3][14][5][7]. An objective of this project was to capture the process of locating errors into a repeatable form and save the quality data states after each execution in order to develop a history of quality issues [5] which will hopefully lead to resolving some of the error sources in the long term [2][3]. Data quality can be approached from a high level or granular perspective; these are referred to respectively as the process or data methodology. This project has used the data-driven methodology on the whole but the DQMDW is moving in the direction of a more process-driven approach. "In general in the long term, process-driven techniques are found to outperform data-driven techniques, since they eliminate the root causes of quality problems. However, from a short term perspective, process redesign can be extremely expensive [Redman 1996][English 1999]. On the contrary, data-driven

strategies are reported to be cost efficient in the short term, but expensive in the long term." [3] Toward the goal of benefiting from both methodologies, a data quality metadata warehouse [8] was developed. There, the results of the profiling, assessments, rules and measurements can be captured. These can then be used to build a data quality history and to maintain a record of the data quality project information and resulting improvements. This can provide a evidence for the continuing support and funding of the data quality project.

The DQMDW supports the TDQM continuous cycle methodology of identifying requirements, measuring the quality, analyzing the results and then data process improvement [4][3][2]. Moving beyond error detection and correction, toward process control and improvement, for long term quality management should include as part of the design supporting quality monitoring with metadata [3].

This work is concerned with the quality of an active database, although many researchers cited due to their interest in metadata were working in the data warehouse area [14]. This work was conducted from the perspective that the use of quality metadata is equally applicable to both the production database as well as the data warehouse. Actually this may be an even more advantageous position in terms of locating error root causes since the data is in a raw state, it has not yet been cleaned and aggregated, which could lead to the removal of the error-cause connection.

4. Visualization

Visualization makes comprehension simpler and helps to convert data to information for the user. In order to take advantage of the data quality history that has been captured a visualization tool for viewing the DQMDW was created. The book "Journey to Data Quality" also suggests first automating quality measurements then displaying these results in a graphical form [4]. The aggregated quality data scores are linked with drill down capabilities to more detailed views [12]. This will provide the user with an overview of the data quality when using it for business decisions and it will motivate the data stewards to sustain the interest and effort in the data integrity project itself [12]. This tool uses the 'Data Quality Assessment' scorecard approach as well to present the quality of the data first as aggregated for a broad overview, then with successively more detailed views eventually displaying the errors directly [8].

IV. DATA QUALITY SETUP: GETTING QUALITY MEASUREMENTS

The university working environment contained several database interaction tools. Toad and SQL are used for direct Oracle access and shell scripting is used in UNIX to run integrity checks. The only expansion needed for the environment was the inclusion of OYSTER and a Data Quality Metadata Warehouse.

1. Environment

A. Database: The Banner database at UALR runs on Oracle vs. 10gR2 on a Solaris UNIX Server.

B. OYSTER: OYSTER is written in Java for platform independence, so setting it up in either the Windows or UNIX environments works well. In this project it was run on a Windows server since the university is in the process of switching the OS for the database server and upgrading Java to a more recent version that is compatible with OYSTER.

C. Scripting: The scripting was done with bash, UNIX shell, scripts in the Solaris UNIX environment where the database is located. They will set up for scheduled execution once a UNIX server with a java version compatible with OYSTER can be obtained.

D. DQMDW: The DQMDW was implemented in the same Oracle database as the Banner database. Since the data used in the project comes from the university, it is subject to Family Educational Rights and Privacy Act, or FERPA. To ensure FERPA compliance a copy of the DQMDW has been populated with synthetic data in order to create a version suitable for project demonstration.

E. Profiling: Profiling was done with the Data Quality Analyzer and Toad tool. Both tools provided column profile information, but the Data Quality Analyzer provided pattern analysis capabilities that aren't available in Toad. Toad's strength is its SQL query tools and general database capabilities that do not exist in the DQ Analyzer.

2. Measurements Taken Previous to this Project

Since 2009 MPOP errors located by the common matching component have been counted and those numbers reported every six months.

3. Measurements Taken During this Project

Student data in the University of Arkansas at Little Rock Banner database is the study population. The subject area is the data quality level before and after a data quality assessment and improvement project. The variables can be operationalized by identifying quality rules and using their results as inputs to calculate the quality of the data to which the rules apply. These rules will be numerous but a few examples are

- 'SSN should have nine digits',
- 'Each student should have only one T-number',
- 'Each student should have address information',
- 'Each address should represent a real place'.

Quality metrics were based on percentages [4][11][10]. The metrics and the quality problems they measured are defined below.

Accuracy – Incorrectly entered SSN numbers, names, addresses, telephone numbers and DOB. This problem is checked with the generic error script, the errors will be counted and the metric applied within the visualizer.

Metric:

$$1 - \frac{Number\ of\ Known\ Errors}{Number\ of\ Elements}$$

Consistency – Primarily PIDM numbers, used to track student entities, are sometimes assigned to one entity multiple times resulting in MPOP (multiple PIDM one person) errors. Other inconsistencies include city / state / zipcode, phone number and SSN formats. This problem is checked with the MPOP and generic error scripts, the errors will be counted and the metric applied within the visualizer.

Metric:

$$1 - \frac{Number\ of\ Consistency\ Violations}{Number\ of\ Data\ Elements}$$

Completeness – Many of the fields contain blank or null values which are both a symptom of past errors as well as a source for new errors. This problem is checked with the generic error script, the errors will be counted and the metric applied within the visualizer.

Metric:

$$1 - \frac{Number\ of\ Blank\ or\ Null\ Elements}{Number\ of\ Data\ Elements}$$

The most important quality dimensions distilled from many prominent authors; accuracy, completeness, consistency, and timeliness, [2][4][8] are measured in this project with the exception of timeliness. Timeliness is a measure that is user dependent and as such varies across the campus. This measurement as well the measurement of believability would have required surveying a large group of users adding to much time and expense to this project. It would also have blurred the project focus from the data perspective [9] and would make automatically repeatable data quality check-ups difficult if not impossible.

V. PROJECT STEPS

1. *Data Quality Assessment*

Generic rules were based on format, interdependencies, optionality or other logical necessities that were not enforced by the database design but were requirements for high quality data. The profiling results indicate problem areas that the scripts can target. Many fields are completely or mostly null. Referential integrity is not enforced for PIDM number so orphaned records can occur. The phone numbers and area codes are not constrained to numeric values. The zip code is also not constrained to numeric values. The SSN records should be unique with the PIDM but this field is not 100% unique. These clues will help to form the structure of the generic errors script.

The MPOP script uses matching rules to locate potential entity resolution errors. The process of deciding which rules are used together is not as obvious as it might seem.

A. Discovering the Data Quality Rules: Intuition guided the original design of the matching rules. Rules that produced few false positives, but numerous results were considered the most fruitful. The six rules listed below were found to return dependably good results so were tested in various combinations to determine the optimal combination of the best rules.

1. PIDM exact
2. FIRST NAME exact, LAST NAME exact, and SSN transposed
3. FIRST NAME exact, DOB exact, and SSN transposed
4. SSN exact and DOB exact
5. FIRST NAME exact and SSN exact
6. FIRST NAME, SSN exact and DOB led

B. Creating Matching Rules: The first rule 'PIDM exact' is used to bring together records that have already been linked by the Banner system. This rule will provide a means for transitive closure to be most effective as well as gather the known sets under one rule that can be discarded later since the unknown group is the object of the search. The ability to locate transposed SSN was not available in the Banner system and proved very useful in locating MPOP errors in combination with other exact match rules.

C. Testing Rules: The results in figure 1 show that rule 4, SSN Exact and DOB exact, is the most productive rule followed by rules 2 and 5. The approximate value for the false positive (≈FP) was obtained by a manual review of the data by the author. The calculations for MPOP percentage rate and FP percentage rate are shown in equation 1 and 2. In order to measure the MPOP rate the Total tuple count is necessary and equals 1,039,013.

$$\approx FP\% = \frac{\approx FP\ Count}{Linked\ Count}\ (100) \qquad (1)$$

Approximate False Positive Percentage

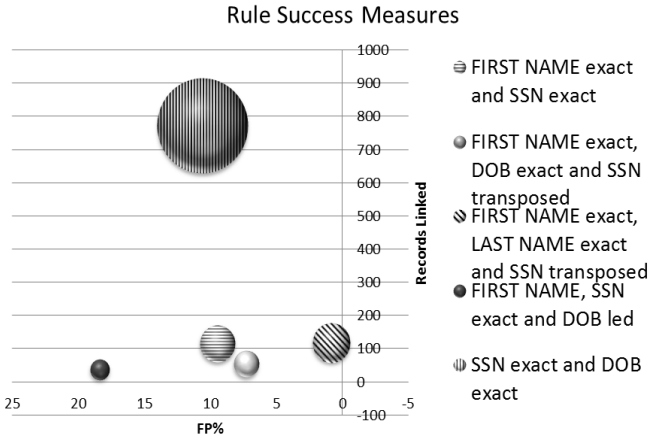


Figure 1: Rule Success
Bubble Size represents the number of correctly returned MPOP

D. Checking Matching Results: The analysis of the results by data experts is recommended by Maydanchik as one method to improve your rules [8]. Each of the proposed MPOP references is to be manually checked by data experts. To check the performance of the rule sets each potential error was checked and marked as a possible false positive if there was not overwhelming evidence that it was a true positive. The exact PIDM rule was not evaluated for false positives since this is asserted knowledge assumed to be asserted knowledge and fully correct, 0% FP rate.

The optimal rule set based on testing was the rule set: 1, 2, 3, and 4 grouped together. Not a surprising result since this set is a combination of the rules with the lowest FP plus the rule with the highest returned MPOP count. Using rule sets developed by combining the experience of the data integrity personnel with trial and error, OYSTER initially located 754 MPOP errors, 237 of which have not yet been verified.

2. *Assessment Results*

A. Oyster:

$$MPOP\ \%\ Rate = \frac{MPOP\ Count}{Total\ Tuple\ Count}\ (100) \qquad (2)$$

MPOP Percentage Rate

Table 1: MPOP Counts

Error Type	Unique PIDM Count	Data Element Total Count	Score
MPOP (Sept 12 2011)	754	1,039,013	0.999274
MPOP (Oct 6 2011)	804	1,039,225	0.999226
MPOP (Oct 31 2011)	903	1,047,292	0.999138
MPOP (Dec 12 2011)	462	1,047,272	0.999559

B. Generic:

Table 2: Generic Error Counts

Error Type	Count	Total Count	Score
duplicatedSSN	651	486,385	0.998662
NoSpbpersP-Entity	15504	486,385	0.968124
numericZip	1038	1,321,579	0.999215
OrphanedSpraddr	3	1,321,579	0.999998
orphanedSpriden	1	1,091,005	0.999999
OrphanedSprtele	1	845,432	0.999999
phoneAreaCdFormatError	143	845,432	0.999831
phoneNumeric	130	845,432	0.999846
stateFormatError	5	1,321,579	0.999996

3. Data Quality Improvement

A. Designing and Populate DQMDW: Maydanchik suggests building a data quality metadata warehouse focused on four core areas of metadata; rule, atomic, general, and aggregate. These sections will be discussed separately and in more detail below.

Rule metadata captures the business and data requirements in the form of data quality rules. This means that all types of quality rules can be found and updated if necessary in one place. Some examples of these entries could include and are certainly not limited to; the rules for record linking will be listed here, rules for data format and optionality, relational integrity rules, attribute dependency rules.

The atomic metadata is the location for staging possible erroneous records and records waiting to be processed. Entity resolution performance is greatly improved by an extract, transform and load (ETL) preprocess to create a dataset for OYSTER that is clean, dimensionally reduced, and integrated. .

Once found these potential MPOP records along with other erroneous records must be kept until they can be corrected. So unlike the data in the other DQMDW tables which may be updated but should not change radically, the data in these tables are transient.

The information gathered in order to create the data quality rules is the source of the general metadata. The data profile and entity relationship information describe the Banner data and are stored in these tables along with the associated table and fields.

The results of quality measurements make up the aggregate metadata. The name and description of scores and the dates of individually recorded scores provide a data quality tracking mechanism. The links to rule and general metadata are also captured both as the source of the score and as areas to be addressed in cases of low scores.

B. Score Design:

$$1 - \left(\frac{Number\ of\ Known\ Errors}{Total\ Number\ of\ Data\ Elements}\right) \quad (3)$$

Accuracy

$$1 - \left(\frac{Number\ of\ Consistency\ Violations}{Total\ Number\ of\ Data\ Elements}\right) \quad (4)$$

Consistency

$$1 - \left(\frac{Number\ of\ Blank\ \frac{and}{or}\ Null}{Total\ Number\ of\ Data\ Elements}\right) \quad (5)$$

Completeness

Table 3: Scores as Related to IQ Problems

Score	Source	Accuracy	Consistency	Completeness
DB Score	Subject Score(Person)	✓	✓	✓
Subject Score (Person)	spbpers errors, spriden errors	✓	✓	✓
SPBPERS Table Score (Person)	NoSpbpersP-Entity, MPOP error		✓	✓
SPRIDEN Table Score (Person)	orphanedSpriden, duplicatedSSN	✓		✓
SPRADDR Table Score (Person)	OrphanedSpraddr, numericZip, stateFormatError		✓	✓
SPRTELE Table Score (Person)	OrphanedSprtele, phoneAreaCdFormat Error, phoneNumeric		✓	✓
MPOP Score	MPOP error		✓	
Single SSN Score	duplicatedSSN	✓		
Addr Score	numericZip, stateFormatError		✓	
Phone score	phoneNumeric, phoneAreaCdFormat Error		✓	
Identity # score	NoSpbpersP-Entity			✓

C. Score Card Design: The logic underling the scorecard was the basis for both the DQMDW design and the visualization design. The domains the scorecard covers are the MPOP and generic data errors from the data perspective. The scorecard has two objectives: to demonstrate the progress of the data quality efforts and to provide feedback for the data quality assessment and improvement process. Together these objectives will fuel the data quality improvement lifecycle by encouraging process evolution and by providing a stream of return on investment evidence. The

metrics in Equations 3, 4 and 5 make up the score decompositions; the summary scores are aggregations of the more granular scores. The atomic data quality information is the list of errors that are the basis of all of the scores.

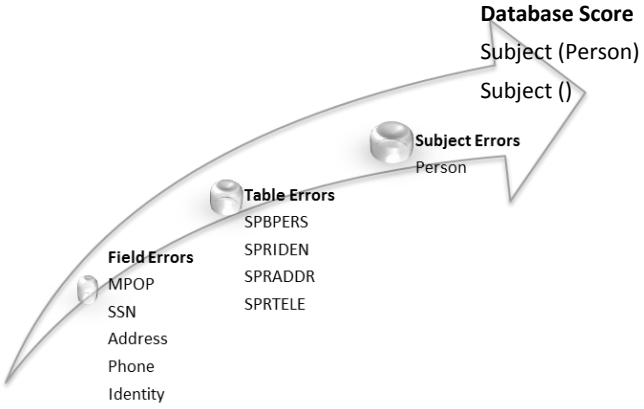


Figure 2 Moving Toward the Database Score
Information granularity decreases with score level

D. Visualizer Design and Build: The dashboard visualization delivers the concepts that the scorecard abstractly presents. The visualization is composed of the subject level scores, table scores, the field scores, and the error listings. This tool prototype was created using Argos, an ERP reporting tool used by the University, that facilitates agile development. This dashboard visualization will not be the final version since it is not portable and cannot be separated from the Argos proprietary product.

VI. ANALYSIS

The script processes have been tested and applied to real data in sections and have not yet been tested as a unit within the script due to hardware and software conflicts. There will, however, be no changes in the database interactions; the processes will be started by an automated process rather than a manual one. It is for this reason that no changes in the results are expected and therefore, the results of these processes shown in the paper should be considered valid. The separated sections have been tested, and the results listed are the results of these efforts.

Since April of 2007, a total of 672 MPOP's were located over six-month intervals. So on average 74.67 MPOP's are identified in each 6-month period or about 13 per month. To compare the OYSTER results to the previous rates we can look at the MPOP counts in table 1 that have been verified. The first MPOP count of 754 is larger than the sum of all of the counts of the last several years. 517 of these have been verified to date leaving 237 with unknown validity, if all of the remaining errors are misidentifications of MPOP errors, which is unlikely, the number of MPOP's identified has still significantly increased from 74.67 per six month period to 517 in one run. It's likely that many of these are not new incoming references but those that have been undetectable in the past and are now apparent due to the improved entity resolution capabilities of OYSTER. This would eventually cause the MPOP counts to decline to only newly entered references which may explain the drop in the December measurement.

The results indicate that this is a sound and productive process. Data errors were not only located, but a process was developed to enable continuous quality monitoring and improvement over time. Quality issues have been identified in the SPRIDEN, SPBPERS, SPRADDR, and SPRTELE tables and are being addressed individually.

These results are useful for both business users as well as IT personnel since they are available graphically. Business users will now have access to student data quality information in an easily understandable visual format when making business decisions based on data. IT personnel can also monitor data quality more efficiently with the assistance of the visual information serving as feedback for their efforts. Finally, this process provides a beginning for an information quality process evolution. As new error types are discovered and added to the scripts or as more subject areas of the database are included these can be added and displayed supporting and encouraging a continued cycle of information quality improvement.

VII. CONCLUSION

Due to the relative youth of our discipline information quality practitioners are faced with the task of maintaining quality without the history of successes and failures that is available in other areas. Other more mature disciplines have studies that can be applied to IQ management and can point us in the right direction, but the task of assessment and improvement can only be tested by performing a study of data quality management in practice. This study coupled with others like it can lay a foundation for data assessment and improvement methodology choices.

The Deming cycle of act, plan, do, check and the TDQM cycle of define, measure, analyse, improve [13] have been the foundation for the design of the process. Starting with a previously existing data integrity project this work has extended the error identification capabilities and created an automated process to save data quality measurements and use them for inputs into the next data quality assessment iteration. Continued executive support and resources for the data quality effort is also more likely due to this visualization aspect of this project.

In the future user defined limits could be added to the visualization tool. When Lee purposed automating the metrics and displaying the results in reports or graphically a more interactive process was described as well as user-defined constraints [4].

ACKNOWLEDGMENT

I would like to thank the UALR Information Technology Services Department and the UALR ERIQ Research Center for their support of this project.

REFERENCES

[1] Arveson, P. (1998). The Deming Cycle. Retrieved March 17, 2012, from BalancedScoreCard.org: http://www.balancedscorecard.org/TheDemingCycle/tabid/112/Default.aspx

[2] Batini, C. C. (2009). Methodologies for Data Quality Assessment and Improvement. *ACM Computing Surveys* 41(3), 16-16.52.

[3] Even, A. A. (2009). Dual Assessment of Data Quality in Customer Databases. *Journal of Data and Information Quality* 1(3), 1-29.

[4] Lee, Y. W. (2006). *Journey to Data Quality*, Cambridge, MA: MIT.

[5] Li, Y. A.-M.-B. (2010). Quality factory and quality notification service in data warehouse. *Proceedings: 3rd*

Workshop on Ph.D. Students in Information and Knowledge Management (pp. 25-32). Toronto, ON, Canada: ACM.

[6] Madnick, S. E. (2009). Overview and Framework for Data and Information Quality Research. *Journal of Data and Information Quality* 1(1), 1-22.

[7] Matthias Jarke, Y. V. (1997). Data Warehouse Quality: A Review of the DWQ Project. *Proceeding: 2nd Conference on Information Quality*. Cambridge, MA.

[8] Maydanchik, A. (2007). *Data Quality Assessment*. Bradley Beach, NJ: Technics Publications.

[9] Mouzhi Ge, M. H. (2007). A Review of Information Quality Research. *Proceedings: 12th International Conference on Information Quality* (pp. 76-91). MIT.

[10] Pipino, L. L. (2002). Data quality assessment. *Communications of the. ACM* 45(4), 211-218.

[11] Talburt, J. R. (2011). *Entity Resolution and Information Quality*. San Francisco, CA: Morgan Kaufmann.

[12] Ward, M. Z. (2011). Quality-aware visual data analysis. *Computational Statistics* 26(4), 567-584.

[13] Yang W. Lee, L. P. (2004). Process-Embedded Data Integrity. *Journal of Database Management* 15(1), 87-103.

[14] Yuriy Verbitskiy, W. Y. (2011). Data Quality Management in a Business Intelligence Environment. *Proceedings of the 16th International Conference on Information Quality (ICIQ-11)*, Adelaide, Australia.

[15] Zhou, Y., Talburt, J. (2011) Entity Identity Information Management. *Proceedings of the 16th International Conference on Information Quality (ICIQ-11)*, pp. 327-341, Adelaide, Australia.

Internal Fixation Evaluation: A Machine Learning Approach

Ray Hashemi[1], Cameron Coates[2], Azita Bahrami[3], Mahmood Bahar[4], Alexander Tyler[5], Nicholas Tyler[6], and William Gibson[1]

[1]Department of Computer Science
[2]Engineering Program
[6]Department of Biology
Armstrong Atlantic University
Savannah, GA 31419, USA

[3]IT Consultation
Savannah, GA

[4]Department of Basic Sciences
Garmsar Branch, Islamic Azad University
Garmsar, Iran

[5]Conway Regional Health System
2302 College Avenue, Conway, AR 72034

Abstract- *A major dilemma currently faced by orthopedic surgeons is whether to: retain or remove an internal fixation from patient's body after the bone rebuilds itself. The difficulty stems from the fact that for both options there are total of ten major side-effects. The goal of this research effort is to generate a set of rules of thumb by which a decision can be rapidly reached. The goal was met by (1) creating organic synthetic data using likelihood measures, (2) Calculating systematically the confidence interval of the risk factors for every side-effect, (3) developing a special type of neural network to pre-process the organic patients' records in reference to the ten side-effects, (4) applying Wilcoxon's statistical model to conclude the rules of thumb, and (5) verifying the rules of thumb using a domain expert which resulted in 87% accuracy.*

Key Words: *Internal fixation, Organic synthetic record, Likelihood measures, Neural network, Risk factors, retaining and removal side-effects.*

1. Introduction

Internal fixation means stabilizing and joining the ends of fractured bones by mechanical devices such as metal plates, pins, rods, wires or screws. An example is shown in Figure 1.

The major question facing orthopedic surgeons is: Is it more beneficial to retain or remove an internal fixation from patient's body after fracture healing has occurred? One may ask why this is a major question. Because both options may result in significant side-effects.

Figure 1: An example of internal fixation borrowed from [1].

In the case of retaining fixation, the major adverse effects are: *Metallosis* (adverse reaction of the soft tissue in the body caused by the presence of excess metal ions due to a nearby metallic implant), *carcinogenesis* (implant presence may turn normal

cells into cancer cells), *Re-fracture leading to complicated revision*, and *Localized Osteopenia* (decreases in bone mineral density in the vicinity of the implant due to contact with the implant itself and/or stress shielding effects of the implant).

In the case of removing the fixation, the major adverse effects are: *Re-fracture risk* (occurrence of a new fracture in the same area at some point in time), *Iotrogenic fracture risk* (additional fracture complications that may occur during removal of the fixation caused by the activities of the operating physician), *Anesthetic complications* (complications resulting from improper application of anesthesia due to human error, equipment failure, or pre-existing patient related factors such as cardiovascular or respiratory disease), *Nerve damage* (adverse effect on nerves that are at or close to the fracture site being affected by regional anesthesia or during the surgery), *Infection* (contamination of the blood due to bacteria), and *Hematoma* (excessive blood leak into tissues where it does not belong caused by the damaged wall of a blood vessel, artery, vein, or capillary) [2, 3, 4, 5].

An experienced domain expert always uses his/her experience to solve a domain-based problem intuitively. In fact, gained experience is manifested in form of rules of thumb that enable the domain expert in his endeavor. The goal of this research effort is to generate a set of rules of thumb by which a decision for a patient with internal fixation can be rapidly reached.

The rest of the paper is organized as follow: The Previous Works is the subject of section 2. Methodology is presented in section 3. Results and Discussion are covered in section 4. Conclusion and Future Research are the subjects of section 5.

2. Previous Works

The closest research activities to the system that is presented and discussed in this paper is the work of Hanson et al. [6]. Henson and his research team reported the results of a survey in which 655 orthopedic surgeons from 65 countries participated. Among other things, the survey tried to get the answer to the question of when removal of an internal fixation is preferred by the surgeons. Fifty-eight percent of the participants are against the removal of internal fixation partially because they do not believe in the severity of the side-effects associated with the retained implants. Forty-eight percent of the surgeons believe there are more risk in removal than retained implants. However, in the case that patient is a child the removal of internal fixation is highly considered. The justification for this exceptional case is that the children have a growing skeleton.

The reader needs to be reminded, the results are extracted from a survey and do not have any clinical trial foundation. In contrast, we try to investigate and build a decision support system that evaluates both removing and retaining of the internal fixation based on the synthetic patients' data using a machine learning approach. To the best of our knowledge there is no report of such a system in literature.

3. Methodology

The goal of this research is met in four steps of (1) Creating synthetic patients' records, (2) Introducing a new neural network for pre-processing the patient's records, (3) Applying Wilcoxon statistical model to the pre-processed records, and (4) obtaining the rules of thumb. Each step is covered in detail in the following four subsections.

3.1 Creating Synthetic Patients' Records

Let $A = \{a_1, \ldots, a_n\}$ and $B = \{b_1, \ldots, b_m\}$ be the set of side-effects for retaining and removing internal fixations, respectively. Comparing the sets A and B is not possible because side-effects in A are different from side-effects in B. To make A and B comparable, we introduce a new set of side-effects that belongs to both retaining and removing of the internal fixations. The new set of side-effects is $S = \{s_1, \ldots, s_{(n+m)}\}$ where, $S = A \cup B$ and $|S| = |A| + |B|$. For the investigation in hand n+m = 10.

Table 1: Patient attributes and categories

Cat.	Age	Weight
1	<16	Underweight
2	16-35	Normal
3	35-60	Over Weight
4	>60	Obese
	Physical Activity Level	**Health Problems**
1	Sedentary	None
2	Moderate	Low
3	Very Active	Medium
4	Elite Athlete	Serious

A patient is modeled by four attributes of *Age*, *Weight*, *Physical activity level*, and *Health problems*. Each attribute has four possible categorical values of 1, 2, 3, and 4, shown in Table 1.

Let us look at two synthetic patients' records that are in the same age group and having the same weight, but Physical Activity Level for the first record is *sedentary* while for the second one is *elite athlete*. Let us assume the value for the forth attribute (Health

problems) of both records is the same and it says the patients have *serious* health problems. Common sense suggests that it is less likely for an elite athlete to have serious health problems. Therefore, the first record looks more *organic* than the second one.

Table 2: Likelihood values for all possible patterns of set (Age, Weight, Physical Activity Level).

Age	Weight	Physical Activity Level			
		1	**2**	**3**	**4**
1	**1**	0.92	**0.02**	0.02	0.0
	2	0.05	0.1	0.84	0.01
	3	0.9	0.1	0.0	0.0
	4	0.99	0.01	0.0	0.0
2	**1**	0.8	0.15	0.05	0.0
	2	0.1	0.3	0.5	0.1
	3	0.78	0.15	0.05	**0.02**
	4	0.98	0.02	0.0	0.0
3	**1**	0.9	0.1	0.0	0.0
	2	0.01	0.64	0.35	0.0
	3	0.85	0.15	0.0	0.0
	4	0.95	0.05	0.0	**0.0**
4	**1**	0.98	**0.02**	0.0	0.0
	2	0.1	0.8	0.1	0.0
	3	0.9	0.1	0.0	0.0
	4	0.95	0.05	0.0	0.0

Another point that needs to be made is that the values for attributes Age and Weight will also correlate with the patient being an elite athlete. For example if the age value for both patients is 4 (i.e. >60) the likelihood of the person being an elite athlete is zero.

To create organic synthetic patient records, the domain expert assigns a *likelihood* value, in the range of [0-1), to every possible combinations of attribute values (*patterns*). This is done systematically by assigning likelihood values to all the possible combinations of attributes of {Age, Weight, and Physical Activity Level} and {Age, Weight, Health Problems}, Tables 2 and 3.

For each Age value, one of the likelihood values is designated as the *threshold* by domain expert and it is shown in bold. Therefore, there are four threshold values for each one of the Tables 2 and 3.

A patient record that is composed of four values is checked against both Tables 2 and 3. If one of the patterns for {Age, Weight, Physical Activity Level} or {Age, Weight, Health Problems}, has the likelihood value less than or equal to the corresponding threshold value, the record is dismissed because it is not organic.

The tables for likelihood values are used by the algorithm ORGANIC for creating organic patient records.

Table 3: Likelihood values for all possible patterns of set (Age, Weight, Health Problems).

Age	Weight	Health Problems			
		1	**2**	**3**	**4**
1	**1**	0.25	0.25	0.25	0.25
	2	0.99	**0.01**	0.0	0.0
	3	0.85	0.15	0.0	0.0
	4	0.55	0.45	0.0	0.0
2	**1**	0.1	0.3	0.3	0.3
	2	0.88	0.02	0.0	0.10
	3	0.5	0.05	0.05	0.4
	4	**0.02**	0.38	0.1	0.5
3	**1**	0.7	0.1	0.1	0.1
	2	0.01	0.03	0.15	0.3
	3	0.85	0.1	0.15	0.1
	4	**0.01**	0.4	0.19	0.4
4	**1**	0.3	0.2	0.2	0.3
	2	0.4	0.1	0.3	0.2
	3	0.2	0.2	0.3	0.3
	4	**0.0**	0.4	0.3	0.3

Algorithm ORGANIC

Given: Table 2, Table 3, set V = {1, 2, 3, 4), four threshold values for Table 2 ($T_{2,1}$, $T_{2,2}$, $T_{2,3}$, $T_{2,4}$), four threshold values for Table 3 ($T_{3,1}$, $T_{3,2}$, $T_{3,3}$, $T_{3,4}$), and four attributes of Age, Weight, Physical Activity Level, and Health Problems.

Objective: Creation of an organic patient record.

Step 1: Randomly generate two values (i, j) from set V and assign them to attributes Age and Weight.

Step 2: Randomly generate a value (k) from set V.
If Table2(i, j, k) $\leq T_{2,i}$
Then go to Step2;
Else Assign k to attribute Physical Activity Level;

Step 3: Randomly generate a value (ℓ) from set V.
If Table3(i, j, ℓ) $\leq T_{3,i}$
Then go to Step3;
Else Assign ℓ to attribute Health Problems;

Step 4: End;

Algorithm ORGANIC recognizes only 113 pattern out of the 256 possible patterns as organic.

To each value in an organic patient's record, 10 side-effects are related. Let v be a possible value for one of the attributes of patient's record and s_i be one of the side-effects. There is a risk factor (probability) involved with s_i in reference to v that is denoted as $P(s_i|v)$. The $P(s_i|v)$ and its confidence interval are calculated using the following procedure.

Procedure: A population, G, of 1000 patients with internal fixation was randomly created. Each patient had four random values (borrowed from set {1, 2, 3, 4}) for the four attributes of Age, Weight,

Physical Activity Level, and Health Problems. Population G included only organic patients' records

Table 4: Risk Factors

Side Effect *	Independent Variables			
	Age			
	1	2	3	4
M	0.7..0.9	0.5..0.7	0.3..0.5	0.2..0.3
C	0..0.3	0.3..0.5	0.5..0.7	0.7..0.9
RC	0.4..0.6	0.4..0.6	0.5..0.7	0.7..0.9
LO	0.2..0.4	0.1..0.2	0.3..0.5	0.2..0.4
RR	0.1..0.3	0.3..0.5	0.1..0.2	0.3..0.5
LF	0.1..0.3	0.05..0.1	0.5..0.7	0.3..0.5
AC	0.3..0.5	0.1..0.2	0.1..0.2	0.3..0.5
ND	0.3..0.5	0.1..0.2	0.1..0.2	0.3..0.5
I	0.4..0.6	0.3..0.5	0.3..0.5	0.4..0.6
H	0.1..0.2	0.2..0.3	0.2..0.3	0.4..0.6
	Weight			
M	0	0	0	0
C	0	0	0	0
RC	0.1..0.2	0.0..0.05	0.3..0.4	0.1..0.2
LO	0.1..0.2	0.0..0.05	0.1..0.2	0.1..0.2
RR	0.3..0.5	0.1..0.2	0.2..0.4	0.3..0.5
LF	0.2..0.4	0.2..0.4	0.3..0.5	0.4..0.7
AC	0.0..0.2	0.0..0.1	0.1..0.3	0.2..0.5
ND	0.2..0.4	0.2..0.4	0.3..0.5	0.3..0.5
I	0.2..0.4	0.2..0.4	0.2..0.4	0.3..0.5
H	0.0..0.2	0.0..0.2	0.1..0.3	0.1..0.3
	Physical Activity Level			
M	0	0	0	0
C	0	0	0	0
RC	0.0..0.05	0.05..1	0.95..0.99	0.3..0.5
LO	0.1..0.2	0.05..0.07	0.1..0.2	0.05..0.1
RR	0.0..0.2	0.2..0.4	0.3..0.5	0.5..0.8
LF	0.2..0.4	0.0..0.2	0.0..0.1	0.0..0.1
AC	0.2..0.4	0.1..0.3	0.0..0.2	0.0..0.2
ND	0.2..0.4	0.1..0.2	0.1..0.2	0.1..0.2
I	0.2..0.4	0.3..0.5	0.2..0.3	0.0..0.2
H	0.1..0.3	0.1.0.4	0.0..0.2	0.0..0.2
	Health Problems			
M	0	0.2..0.3	0.05..0.1	0.1..0.2
C	0	0	0	0
RC	0	0.3..0.4	0.1..0.2	0.2..0.3
LO	0.0..0.1	0.05..0.1	0.1..0.2	0.1..0.2
RR	0	0.3..0.4	0.1..0.2	0.2..0.3
LF	0	0.2..0.3	0.05..0.1	0.1..0.2
AC	0.0..0.2	0.05..0.1	0.1..0.2	0.2..0.3
ND	0	0	0	0
I	0.01..0.04	0.05..0.1	0.05..0.1	0.1..0.2
H	0.1..0.2	0.05..0.1	0.2..0.3	0.3..0.4

* **M**: Metallosis, **C**:carcinogenesi, **RC**: Re-fracture leading to complicated revision, **LO**: Localized Osteopenia, **RR**: Re-fracture risk, **LF**: Lotrogenic fracture risk, **AC**: Anesthetic complications, **ND**: Nerve Damage, **I**:nfection, and **H**: Hematoma.

.

In addition, one side-effect, s_i, was added to G. Value of s_i for a given record was randomly assigned and it was either 0 or 1. Zero means the patient did not suffer from s_i and "1" means otherwise.

Let K be the number of patients in G suffering from s_i and let L be the number of patients within K with value v for a designated attribute. In addition, Let M be the number of patients in G with value v for the same attribute. The following probabilities can be calculated for the population: $P(s_i) = K/|G|$, $P(v|s_i) = L/K$ and $P(v) = M/|G|$. Out of these probability one can calculate $P(s_i|v)$ using formula (1):

$$P(s_i|v) = \frac{P(v|s_i).P(s_i)}{P(v)} \qquad (1)$$

The process was repeated 10 times and each time a new randomly generated population was created and $P(s_i|v)$ was calculated for the new G.

Let μ be the mean for $P(s_i|v)$ and let the ten values of $P(s_i|v)$ be denoted as x_j, j = 1 to 10 . The objective here is to establish a confidence interval for μ. The estimated mean is $\bar{X} = \frac{1}{10}\sum_1^{10} x_j$ and the estimated variance of the probabilities is $S^2 = \frac{1}{9}\sum_1^{10}(x_j - \bar{X})^2$. The (1-α)% confidence interval for μ is given by formula (2):

$$\bar{X} - t_{u,\alpha/2}\ \frac{S}{\sqrt{10}} \le \mu \le \bar{X} + t_{u,\alpha/2}\ \frac{S}{\sqrt{10}} \qquad (2)$$

where $t_{u,\alpha/2}$ is the t-value with u = 9 degrees of freedom from the student t probability distribution, leaving an area of α/2 to the right.

End of procedure.

We used α = 0.05 for the calculation of the confidence intervals. For s_i = *Metallosis*, the obtained interval value for v = 1 of the attribute Age is [0.7, 0.9]. This means, P(*Metallosis* | 1) for 95% of the patient populations falls in the range of [0.7-0.9].

The above procedure was repeated to obtain the interval of the risk factor for all possible values of the four attributes and for all 10 side-effects. The results are shown in Table 4. Some of the intervals in the table have zero as a value and it means there is no risk factor for the corresponding side-effect. The confidence intervals of zeros are not produced by the procedure and they are suggested by the domain expert.

3.2 A Neural Network for Pre-Processing the Patients' Records

Each patient's record has four risk factors for every side-effect which their sum represents the *strength* of the side-effect for the record. In addition, each side-effect influences decisions for both retaining and removing of the internal fixation. This influence

is represented by a *weight* assigned to each side-effect by the expert. The side-effects' weights are shown in Table 5.

Table 5: The weights for the side-effects' weights.

Retaining		Removing	
Side Effect	**Influence Weight**	**Side Effect**	**Influence Weight**
Metallosis (M)	1.2	Re-Fracture within 18 months (R)	1.2
carcinogenesis (C)	1	Lotrogenic fracture (L)	1
Re-Fracture and Revision (r)	1.6	Anesthetic Complications (A)	1
Osteopenia (O)	1	Nerve Damage (N)	0.3
		Infection (I)	0.5
		Hematoma (H)	0.4

A new neural network (Figure 2) is developed that is able to (1) calculate the strength of each side-effect for every patient's record (2) treat all the side-effects pertaining to retaining decision as one entity and all the side-effects related to the removing decision as another entity, and (3) deliver a quantitative influence for each entity on their corresponding decisions. The output of the neural net is directly used by the Wilcoxon model.

The neural network is made up of three layers. The first layer, input layer, accepts a patient's record. Considering the fact that each patient's record has four values, the number of nodes in the first layer is four.

The i-th node (for i = 1 to 4) of the input layer is made up of a look-up table with four columns and ten rows. Each column represents risk factors for one of the four possible values of the i-th attribute. Each node uses its input as column index and then the ten values of the selected column serve as the output of the node. The weight matrix, W, for all the connections between the first and the second layer are initialized with value of one.

The second layer, hidden layer, has ten nodes representing the ten side-effects. Each node has four inputs from the first layer and produces one output using the following formula:

$$\sigma_j = \sum_{i=1}^{4} Input_{i,j} * w_{i,j} \quad (For\ j = 1\ to\ 10) \qquad (1)$$

where, $Input_{i,j}$ is the input from $node_i$ of the input layer to $node_j$ of the hidden layer and $w_{i,j}$ is the weight for the connections between $node_i$ of the input layer and $node_j$ of the hidden layer.

The third layer, output layer, has only two nodes of A and B. Node A receives only the output of the first four nodes of the hidden layer and the node B receives the output of the last six nodes of the hidden layer. The output for the nodes A and B are calculated using formulas 2 and 3.

$$x = \sum_{j=1}^{4} \sigma_j * u_j \quad (2)$$
$$y = \sum_{j=5}^{10} \sigma_j * u_j \quad (3)$$

where, σ_j is the input from $node_i$ of the hidden layer to the corresponding node in the output layer and u_j is the weight for the connections between $node_j$ of the hidden layer and the corresponding node in the output layer and its value is borrowed from Table 5.

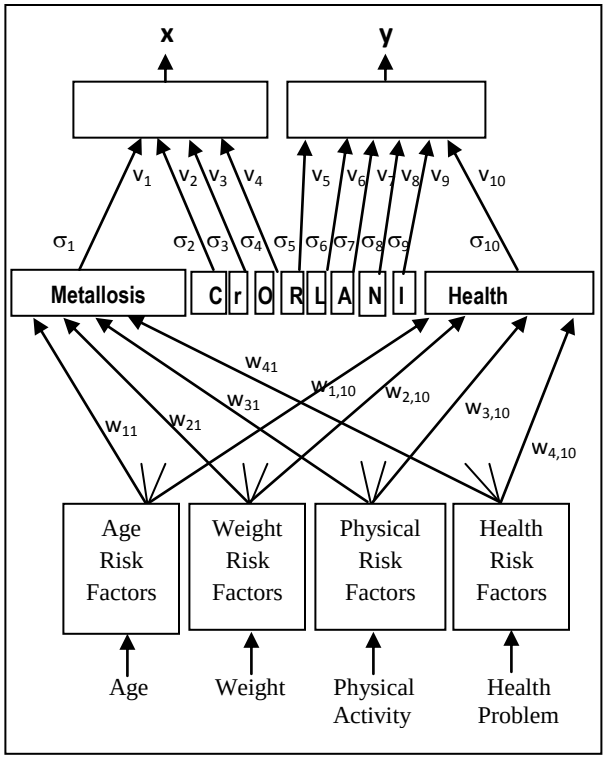


Figure 2: The neural Network

The output of the first and the second nodes in the third layer are referred to as x and y and represent the final outcome of a pre-processed patient's record.

.

3.3 The Wilcoxon Statistical Model

Let S1 and S2 be two samples of a population (S1 and S2 are not necessarily distinct) and E1 and E2 be two experimental events (either of the two events can be null but not both of them). Let S1 be exposed to E1 and S2 to E2. In addition, let the differences observed between S1 and S2 after exposure be D. If D is not significant, the null hypothesis (that E1 and E2 did not have lasting effects on S1 and S2) is rejected and the

alternative hypothesis that E1 or E2 has a lasting effect on its corresponding sample is accepted.

As an example, two groups of patients (S1 and S2) who have the same sickness have been selected. One group (S1) is treated by a new drug (E1) but the second group (S2) is not treated at all (E2= ∅). If the differences observed between the two groups (D) is not significant then the drug is not effective on the sickness; otherwise it is.

For the problem that in hand, S1 and S2 are the same (the same sample of population). E1 and E2 are all the risk factors for side-effects of *retaining* and *removing* internal fixation, respectively. The null hypothesis (H_0) is that the there is no preference in either retaining or removing the internal fixation and the alternative hypothesis (H_1) is that there is a preference.

The significance of D may be determined by Wilcoxon [7, 8] statistics, or paired t test [9]. The former one is chosen for two reasons: (1) it is used for population with non-Gaussian distribution and (2) it adapts to arbitrary sample size.

The Wilcoxon model is applied to either reject or accept the null hypothesis when the null hypothesis is rejected, then H_1 suggests that one of the decisions (retaining or removing) is preferred. The following algorithm, DECISION, is used to determine the preferred one.

The algorithm DECISION works with a set of randomly generated population, S, for a given pattern. In this population, the randomly generated risk factor for each side-effect is in the range prescribed by Table 4 for the pattern. The difference between side-effects for retaining and removing the internal fixation of the population S is calculated using formula (4).

$$D = \sum_{i=1}^{|S|} E_{1,i} - \sum_{i=1}^{|S|} E_{2,i} \qquad (4)$$

Based on the significance of D, the decision outcome for a valid pattern is: *No-Preference*, *Retaining*, or *Removing*.

Algorithm DECISION

Given: E1 and E2 (All the risk factors for side-effects of *retaining* and *removing* internal fixation). A null Hypothesis (H_0) that says no preference between E1 and E2. A population S that includes randomly generated patients' records for a given valid pattern. Wilcox critical value table and the neural network of Figure 2.

Objective: Evaluate the internal fixation.

Step 1: Pre-process S using the neural network; X=$\sum_1^{|S|} x_i$ and Y=$\sum_1^{|S|} y_i$;

Step 2: Using the Wilcoxon nomenclature, X and Y are considered total absolute value of positive and negative ranks, respectively;

Step 4: R = Min(X, Y); and the sign of R is borrowed from D.

Step 4: Use Wilcoxon critical value table to get the sum of total rank (ρ) using number of patients (df) and confidence level α=0.05.

Step 5: If (ρ >R)
Then /*H0 is true */
Decision← No-Preference;
Else hypothesis is false;
If(R is positive)
Then Decision← Retaining;
Else Decision← Removing;

Step 6: End;

Step 5 needs further explanation. In the case of null hypothesis rejection, if R=X, it means the rank value for X (i.e. retaining) was smaller than rank value for Y. Therefore, total risk probability for retention is less than the total risk probability for removal. Thus, decision is for "retaining". If R =Y using the same reasoning, decision is for "removing".

3.4. Rules of Thumb

Each pattern with a decision can be presented as an *if-then* rule. For example, if for the pattern PAT = "1223" the decision is *retaining*, then the following rule with four conditions can be generated:

If (Age = 1) ^ (Weight = 2) ^
(Physical Activity Level = 2) ^
(Health Problems = 3)
then Decision = Retaining.

Therefore, we use the terms attribute and condition interchangeably.

Table 6: Patterns with mixture of decisions

#	Pattern	Decision	#	Pattern	Decision
1	1 2 1 3	Remove	6	1 3 3 3	Remove
2	2 1 4 1	Retain	7	2 4 4 3	Remove
3	3 3 3 3	Retain	8	2 3 4 1	Retain
4	4 3 3 2	Retain	9	1 2 2 4	Retain
5	2 1 2 3	Remove	10	3 2 1 3	Retain

Rules of thumb are the generalization of retaining and removing rules that are compact and easy to remember. We generalize the rules using a modified Dropping Condition Approach [10]. In this approach, a minimum subset of conditions (attributes) in a given rule is kept such that the values for the subset of attributes can be found only in the removing rules, for example, and not in any of the retaining rules.

To provide an example of the generalization approach, let the patterns of Table 6 have a mixture of

retaining and removing decisions. The four values in each pattern represent the attributes Age, Weight, Physical Activity Level, and Health Problems, respectively. Values for any of the attributes cannot exclusively identify patterns for one of the decisions. However, values for Age and Health Problems, collectively, can identify all the patterns of the removing decision and none of the retaining decision. One may conclude that if combination values for Age and Health Problems can represent the patterns of removing decision by inclusion, it can also represent the patterns of retaining rules by exclusion. As a result, the following general rule can represent both removing and retaining rules:

If ((Age = 1)$^\vee$ (Age = 2)) $\wedge$
(Health Problems = 3)
Then Decision = Removing;
Else Decision = Retaining;

The above rule is compact and easy to remember.

4. Results and Discussion

There is a total of 256 patterns (four attributes and four possible values for each attribute) from which 113 of them are valid patterns (using algorithm ORGANIC).

Let PAT = "1223" be a valid pattern. Each value in this pattern has a range of risk factors for the ten side-effects. Therefore, one can generate M number of patients' records for which the pattern is the same but the risk factors for each value of the pattern may be different. Analysis of the M records using Algorithm Decision produces one of the following three decisions for PAT: *no-preference*, *retaining*, and *removing*.

Results revealed that no patterns has the decision no-preference, 92 patterns have the decision retaining, and 21 patterns have the decision removing. The generalization of the rules generated the following rule of thumb:

If (Weight $\neq$ 4) $\wedge$
((Physical Activity Level = 3)$^\vee$
(Physical Activity Level = 4))
Then Decision = Removing;
Else Decision = Retaining;

To validate the above rule of thumb, (1) decisions generated for all 113 valid patterns shared with domain expert and 87% of the decisions were confirmed by the expert.

5. Conclusion and Future Research

An extra effort has been dedicated to the creation of the organic patient records. Use of likelihood measures and confidence interval to determine the range of risk factors for each side-effect are crucial in support of having randomly generated organic patients' records. The outcome of the verification of the rule of thumb by the domain expert is a good indicator for: (1) the quality of the organic patient's records and (2) viability of the presented methodology.

As future research, the collection of data from the real patients and assembling a team of domain experts for verification of the rules of thumb generated by the presented methodology are in progress.

6. References

1. http://en.wikipedia.org/wiki/Internal_fixation, retrieved in November 2011.

2. SB Keel, KA Jaffe, G Petur Nielsen, AE Rosenberg, "Orthopaedic implant related sarcoma: A study of twelve cases," Mod Pathol, 14(10), 969-977, 2001.

3. LB Signorello, W Ye, JP Fryzek, L. Lipworth, JF Fraumeni Jr., WJ Blot, JK McLaughlin,O. Nyrén, "Nationwide study of cancer risk among hip replacement patients in Sweden," J Natl Cancer Inst, 93(18), 1405-1410, 2001.

4. ML Busam, RJ Esther, WT Obremskey, "Hardware Removal: Indications and Expectations," J Am Acad Orthop Surg, 14(2),113-120, 2006.

5. W. Jamil, M, Allami, MZ Choudhury, C. Mann, T. Bagga, A. Roberts, "Do orthopaedic surgeons need a policy on the removal of metalwork? A descriptive national survey of practicing surgeons in the United Kingdom", Injury, Int. J. Care Injured, 2008, 39(3), 362-367, 2008.

6. B. Hanson, ML CVD Werken, and D. Stengel, "Surgeons' beliefs and perceptions about removal of orthopaedic implants", BMC Musculoskeletal Disorders 2008, 9:73.

7. M. Hollander and D. A. Wolfe, Nonparametric Statistical Methods, 2nd Ed., A Wiley Interscience Publication. 1999.

8. G.W. Corder and D.I. Foreman, Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach, New Jersey: Wiley, 2009.

9. H. Motulsky, Intuitive Biostatistics, Oxford University Press Inc, 1995.

10. R. Hashemi, A. Tyler, A. Bahrami, "Use of Rough Sets as a Data Mining Tool for Experimental Bio-Data", "Computational Intelligence in Biomedicine and Bioinformatics: Current Trends and Applicationjs", Springer-Verlag Publisher, June 2008, pp. 69-91.

SESSION

NOVEL APPLICATIONS: INFORMATION AND KNOWLEDGE ENGINEERING + DATA AND INFORMATION MINING

Chair(s)

Prof. Hamid R. Arabnia

Mining Frequent Patterns Based on Data Characteristics

Lan Vu, Gita Alaghband, Senior *Member, IEEE*
Department of Computer Science and Engineering, University of Colorado Denver, Denver, CO, USA
{lan.vu, gita.alaghband}@ucdenver.edu

Abstract - *Frequent pattern mining is crucial part of association rule mining and other data mining tasks with many practical applications. Current popular algorithms for frequent pattern mining perform differently: some are good for dense databases while the others are ideal for sparse ones. In our previous research, we developed a new frequent pattern mining algorithm named FEM that runs fast on both sparse and dense databases. FEM combines the mining strategies of FP-growth and Eclat and given a user-specified threshold it adapts its mining behaviors to the data characteristics to efficiently find all short and long patterns from different database types. However, for best performance of FEM, an appropriate threshold value used to control the switching between its two mining tasks need to be selected by the user. In this paper, we present DFEM, an improved algorithm of FEM that automatically adopts a runtime dynamic threshold to better fit to the characteristics of the databases. The experimental results show that DFEM outperforms FEM and other popular frequent pattern mining algorithms including Apriori, Eclat, FP-growth on both sparse and dense databases.*

Keywords: data mining, frequent pattern mining, association rule mining, frequent itemset, transactional database.

1 Introduction

Frequent pattern mining is a fundamental task in data mining which is used to find many types of relationships among variables in large databases such as associations [1], correlations [2], causality [3], sequential patterns [4], episodes [5] and partial periodicity [6]. Moreover, it helps in data indexing, classification, clustering, and other data mining tasks as well [7]. Thus, frequent pattern mining has become a focused research with numerous practical applications including consumer market-basket analysis, web mining, similarity search of complex structured data, network intrusion detection and many others [7], [8].

The frequent pattern mining problem aims to search for groups of itemsets, subsequences, or substructures that co-occur in a database with their frequency no less than a user-specified minimum support threshold. For example, a set of items (itemset), such as milk and bread that appear frequently together in a database is a frequent itemset or frequent pattern. In a typical transactional database, the number of distinct single items and their combinations are usually very large. For a small minimum support threshold, the number of generated itemsets can be extremely large. Hence, it is a great challenge to design algorithms for mining frequent patterns that scale with memory size and run in reasonable time [9]. For this reason, many methods for mining frequent patterns have been developed in last two decades [7], e.g Apriori [1], Eclat [10], FP-growth [11], direct hashing and pruning (DHP) [12], sampling technique [13], dynamic itemset counting (DIC) [14], AIM [15], mining using diffsets [16], technique to reduce the FP-tree traversal time [18], H-mine [19] and nonordfp [20]. Among those, Apriori, Eclat and FP-growth are the most well-known and widely used. The efficiency of these three algorithms has been demonstrated by many researchers. However, performance of these algorithms is significantly different for different database types. For example, Eclat works well on dense databases [15], [21], [22] while FP-growth runs faster on sparse ones [11], [17], [18], [19], [20], [23]. Therefore, it is difficult to select a suitable algorithm for specific applications. In addition, for commercial database systems like Oracle RDBMS, MS. SQL Server and IBM DBS2 and statistical software like R, SAS and SPSS Clementine which support data mining tasks [24], [25], [26], the characteristics of databases vary depending on their real applications. For the efficiency of these systems and applications of frequent pattern mining, it is essential to have algorithms that work efficiently for most database types.

Contribution: In this paper, we present DFEM, a new frequent pattern mining algorithm that combines the techniques of FP-growth and Eclat and depends on the data characteristics to select an appropriate technique for each subparts of the database to efficiently discover all long and short frequent patterns from sparse and dense databases. DFEM is based on and a major improvement of FEM, our previously developed algorithm [27]. Unlike FEM, it does not need a pre-determined, user specified threshold on when to switch between the two mining strategies. It automatically finds a runtime dynamic threshold to adjust its search behavior based on the characteristics of databases to improve the mining performance, especially when minimum support threshold is low.

In a brief overview, DFEM uses FP-tree to compact the database in memory and recursively mine the frequent patterns from this data structure like the FP-growth approach. During the mining process, if a conditional pattern base [11] is small enough to be better mined using vertical data structure, it is converted to TID bit vectors and

a weight vector. The algorithm then switches from mining FP-trees using FP-growth to mining TID bit vectors using an approach improved from Eclat. The switching decision is based on a threshold *K* whose value is measured at runtime from data being processed.

We benchmarked DFEM and six other algorithms including Apriori [1], Eclat [10], FP-growth [11], FP-growth* [18] AIM2 [21], FEM [27]. The experimental results show that DFEM outperforms all of these algorithms for many real sparse and dense datasets. The performance merit of DFEM is obtained by efficiently distributing some subparts of the database to be mined by its FP-tree mining task and the other ones by its TID bit vector mining task.

The rest of this paper is organized as follows. Section 2 provides the background knowledge. The DFEM algorithm is presented in Section 3. Section 4 introduces the method of finding the threshold *K*. Experiments and discussion are presented in Section 5. The final section summarizes our study and points out some future research directions.

2 Background

In this section, we describe the frequent pattern mining problem and revisit FP-growth and Eclat, the two mining methods from which our proposed algorithm is derived, to analyze their features, strengths and weaknesses.

2.1 Frequent pattern mining problem

The frequent pattern mining problem can be stated as follows: Let $I = \{i_1, i_2, \ldots, i_n\}$ be the set of all distinct items in the transactional database *D*. The *support* of an *itemset* α (a set of items) is the percentage of transactions containing α in D. A *k-itemset* α, which consists of *k* items from *I*, is frequent if α*'s support* is no fewer than *minsup*, where *minsup* is a user-specified minimum support threshold. Given a database *D* and a *minsup*, the problem statement is to find the complete set of frequent itemsets in *D*. We use the two terms *pattern* and *itemset* as well as *database* and *dataset* interchangeably in this paper. For example, given the dataset in Table 1 and *minsup*=30%, the frequent 1-itemsets include *a, b, c, d* and *e* while *f* and *g* are infrequent because their support is only 22%. Similarly, *ab, ac, ad, ae, bc, bd* are frequent 2-itemsets and *abc* is the only frequent 3-itemset.

TABLE 1
A DATASET WITH MINIMUM SUPPORT THRESHOLD = 30%

TID	Items	Sorted frequent items
1	b,d,a	a,b,d
2	c,b,d	b,c,d
3	c,d,a,e	a,c,d,e
4	d,a,e	a,d,e
5	c,b,a	a,b,c
6	c,b,a	a,b,c
7	f,g	
8	b,d,a	a,b,d
9	c,b,a,e,f,g	a,b,c,e

The exponential search space of itemsets makes finding frequent patterns a nontrivial task. Hence, many scientists have been working on developing efficient and scalable algorithms for frequent pattern mining. Most proposed methods are heuristics to trim down the search space in order to make the solution of this problem feasible. Apriori, Eclat and FP-growth are the most popular and widely used algorithms for mining frequent patterns.

2.2 The Eclat algorithm

Eclat is a popular frequent pattern mining algorithm developed by Zaki *et al.* [10]. It deploys a depth-first search strategy and requires only one database scan. To find frequent k-itemsets, the TID-lists of frequent (k-1)-itemsets, the vertical data layout, are intersected and the frequencies of their resulting lists are computed. The support of an itemset can be easily computed without multiple database scans as needed in the Apriori approach. Hence, the I/O cost is reduced considerably. Eclat has been shown to be one of the best algorithms for long patterns and/or dense databases [15], [21], [22]. Although this method takes advantage of the candidate generation approach, its depth-first search order may require more infrequent itemsets generated and tested than Apriori does. As a result, Eclat's efficiency reduces for sparse databases with short patterns where most itemsets are infrequent.

2.3 The FP-growth algorithm

FP-growth another well-known algorithm proposed by Han *et al.* [11] for frequent pattern mining. It adopts a data structure called FP-tree to compress and store data in memory. It then constructs conditional FP-trees and recursively mines these trees to find all frequent patterns without the requirement of generating a large number of frequent pattern candidates. FP-growth was shown to outperform previously developed methods including Eclat and Apriori [11], [17], [18], [19], [20]. For some dense databases or mining with low minimum support, the number of frequent patterns is very large. For each frequent k-itemset, FP-growth creates a conditional FP-tree used to find the frequent (k+1)-itemsets. Thus, the cost of generating a large number of FP-trees results in the degradation of performance. In such cases, FP-growth does not work as well as Eclat [15], [21], [22].

3 The DFEM algorithm

3.1 Mining frequent patterns based on data characteristics

Studying many real databases and their characteristics [28], [30], we found that most consist of a group of items occurring much more frequently than the others. The ratio of items in this group is smaller for sparse databases and larger for dense ones. Because of their high *support*s, these items appear in most transactions as well as most frequent patterns discovered from a database. If we remove the less frequent items from all transactions of a database, the

remaining database will have the characteristics of a dense database and the removed part will have the characteristics of a sparse one. Moreover, the FP-tree of the FP-growth algorithm is constructed with the nodes of the most frequent items on the top because items are added into FP-tree in the frequency descending order. During the mining process of FP-growth, conditional FP-trees [11] are recursively constructed from parent trees. Our study shows that the size of these trees will reduce to a level where the conditional FP-trees mostly consist of items with high frequencies and have the characteristics of small dense databases [27]. According to the review of Eclat and FP-growth in Section 2, we propose a mining method that applies Eclat's strategy for the dense part and FP-growth's strategy for the sparse part will be more efficient than either Eclat or FP-growth alone.

In our previous study [27], we developed FEM, a frequent pattern mining algorithm based on above observation. FEM combines the mining techniques of Eclat and FP-growth and applies suitable mining strategy for each subparts of database. It has been shown to work better than many popular algorithms including Eclat and FP-growth on both sparse and dense databases by automatically distributing the workload to its two mining tasks where the dense data parts are handled by the mining task inherited and improved from Eclat and the sparse ones are mined using another mining task similar to FP-growth. FEM does this by applying a predefined threshold *K* reflecting the data specific characteristics to decide when to switch between the two mining tasks.

In this paper, we present, DFEM, a dynamic version of FEM that adopts a dynamic threshold *K* whose value is calculated at runtime from data being processed to better fit to data characteristics. This approach relieves the user from finding the appropriate value of *K*. DFEM performs better than FEM and other popular algorithms, especially when minimum support threshold (*minsup*) is low.

3.2 Overview of the DFEM algorithm

DFEM combines the techniques used in the FP-growth and Eclat algorithms. It uses FP-tree to store the compact database in memory and recursively mines the frequent patterns from this data structure similar to FP-growth. During its mining process, DFEM automatically switches between mining FP-trees using FP-growth to mining TID bit vectors using an approach improved from Eclat. The switching decision is based on a runtime dynamic threshold K measured from data. DFEM consists of the three main tasks:

FP-tree construction: Database is scanned for the first time to find the frequent items and create the header table. A second database scan is conducted to get frequent items of each transaction and insert these items into the FP-tree in their frequency descending order. Figure 1 demonstrates the FP-tree generated from the dataset in Table 1.

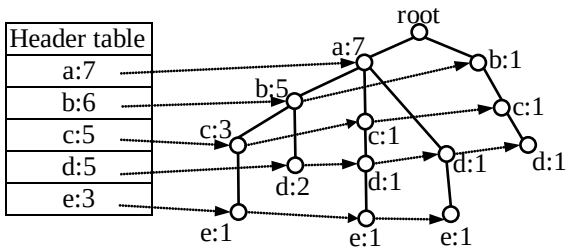


Fig. 1 FP-tree constructed from the dataset in Table 1

FP-Tree mining: This task uses the mining solution of FP-growth where frequent patterns are generated from the conditional FP-trees constructed recursively [11]. However, DFEM is different from FP-growth because only a subset of conditional pattern bases is used to create the conditional FP-trees. The conditional pattern bases whose size are smaller or equal to a threshold *K* will be transformed into TID bit vectors and a weight vector (see Section 3.4) and mining process switches to the *TID bit vector mining* task. K is computed at runtime using a measurement on data being processed as described in Section 4.

TID bit vector mining: This task obtains the TID bit vectors and continues searching for frequent patterns recursively by logical ANDing these bit vectors. The new patterns are constructed by concatenating the suffix pattern of previous steps with the newly generated frequent patterns. This mining task is inspired by Eclat's mining strategy [10]. However, TID bit vectors are used instead of the TID-lists due to their efficiency in computational time and memory consumption [8].

3.3 Algorithmic description

DFEM consists of three main sub algorithms: DFEM, MineFPTree and MineBitVector.

DFEM algorithm: This algorithm (Figure 2) builds the FP-tree, initializes the threshold K using *UpdateK* method (Figure 6) and invokes *MineFPTree* (Figure 3).

DFEM algorithm
Input: Transactional database *D* and *minsup*
Output: Complete set of frequent patterns
1: Scan *D* once to find all frequent items
2: Scan *D* a second time to construct the FP-tree *T*
3: *items* = the number of frequent items in *D*
4: *trans* = the number of transactions in *D*
5: Call UpdateK(*items*,*trans*)
6: Call MineFPTree(*T*,∅,*minsup*)

Fig. 2 DFEM algorithm

MineFPTree algorithm: This algorithm (Figure 3) executes the *FP-tree mining* task. If the condition at Line 13 is not satisfied, *MineBitVector* (Figure 4) will be invoked. Otherwise, *MineFPTree* will recursively mine work in recursive manner to generate frequent patterns from FP-tree data structure. Line 5 and 12 are used to update threshold K using the method described in Section 4.

MineFPTree algorithm

Input: Conditional FP-Tree *T*, *suffix*, *minsup*
Output: Set of frequent patterns
1: **If** FP-tree *T* contains a single path *P*
2: { **For each** combination *x* of the items in *P*
3: { Output $\beta = x \cup suffix$ }
4: *n* = the number of outputs β
5: Call UpdateK(*n*,*1*) }
6: **Else**
7: { **For each** item α in the header table of FP-tree *T*
8: { Output $\beta = \alpha \cup suffix$
9: Construct α's conditional pattern base *C*
10: *n* = the number of items in C
11: *size* = the number of nodes in the linked list of α
12: Call UpdateK(*n*, *size*)
13: **If** *size* > *K* Then
14: { Construct α's conditional FP-tree *T'*
15: Call MineFPTree(*T'*,β,*minsup*) }
16: **Else**
17: { Transform *C* into TID bit vectors *V*
18: and weight vector *w*
19: Call MineBitVector(*V*,*w*,β,*minsup*) }
20: } }

Fig. 3 MineFPTree algorithm

MineBitVector algorithm: This algorithm (Figure 4) executes the *TID bit vector mining* task as described in Section 3.2.

MineBitVector algorithm

Input: Bit vectors *V*, weight vector *w*, *suffix*, *minsup*
Output: Set of frequent patterns
1: Sort *V* in support-descending order of their items
2: **For** each vector v_i in *V*
3: { Output β = item of $v_i \cup suffix$
4: **For** each vector v_j in *V* with $j < i$
5: { $u_j = v_i$ AND v_j
6: sup_j = support of u_j computed using *w*
7: If $sup_j \geq minsup$ Then add u_j into *U*
8: }
9: **If** all u_j in *U* are identical to v_i
10: **Then For each** combination *x* of the items in *U*
11: Output $\beta' = x \cup \beta$
12: **Else If** *U* is not empty
13: **Then** Call MineBitVector(*U*,*w*,β,*minsup*)
14: }

Fig. 4 MineBitVector algorithm

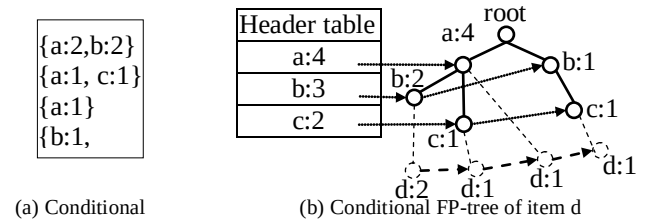
3.4 Transforming a conditional pattern base into TID bit bectors and a weight vector

This is an important step to enable the mining process using TID bit vectors. During the *FP-tree mining* stage, thousands or even millions of conditional pattern bases are processed. However, only those whose sizes are considerably small are transformed into TID bit vectors and weight vector. The size of a conditional pattern base is the number of sets in that base [11]. We use the number of nodes in the linked list of the item of the currently processed FP-tree to decide whether to switch from *FP-tree mining* to *TID bit vector mining* because the number of sets in a conditional pattern base is bounded by and equivalent to this number. The transformation is executed in following steps:

1. Given a conditional pattern base with sets of n items, create n bit vectors whose size are equal to the number of sets and initialized to zero.
2. For each item in the i^{th} set, set the i^{th} bit of its vector to one.
3. Repeat step 2 for every available sets in the conditional pattern base.

Furthermore, each set in a conditional pattern base has a frequency value indicating the number of its occurrence. We combine all frequency values into a weight vector which is used to compute the *support* of items or itemsets.

For example, Figure 5-a presents the conditional pattern base of item d of the FP-tree in Figure 1. If its size does not satisfy the condition for *TID bit vector mining*, this conditional pattern base is used to create the conditional FP-tree (Figure 5-b). Otherwise, it will be transformed into the TID bit vectors (Figure 5-c) and the weight vector (Figure 5-d) by applying above transformative steps.



(a) Conditional pattern base of item d

(b) Conditional FP-tree of item d nodes in dot lines are not part of this tree

a	b	c
1	1	0
1	0	1
1	0	0
0	1	1

(c) TID bit vectors

w
2
1
1
1

(d) Weight vector

Fig. 5 TID Bit vectors and weight vector transformed from conditional pattern base of item d

4 Computing a dynamic K in DFEM

DFEM is developed to automatically determine a good threshold *K* that controls the switching between two mining tasks. This is the main difference between DFEM and FEM. In FEM, the value of *K* is predefined and users need to manually specify a fixed value of *K* for FEM to perform well on a certain database [27]. In contrast, DFEM estimates the near-optimal dynamic value of K at runtime.

4.1 Studying impact of varying K on FEM

We arrive with our approach in DFEM by studying and analyzing impact of varying *K* on FEM. When *K* increases starting from 0, the mining time of FEM and the number of frequent patterns found solely by the *FP-tree mining* task reduce because more conditional pattern bases satisfy the switching condition. FEM will perform best when *K* is equal to a value that presents the best cut off between sparse and dense portions of the database handled by the *FP-tree mining* and the *TID bit vector mining* tasks respectively. For K_i larger than this best *K*, a portion of sparse data is shifted to *TID bit vector mining* which reduces the benefits of *FP-tree mining* and results in unchanged or

increasing the running time of FEM. Table 2 presents the impact of varying *K* on FEM for the Kosarak dataset [28]

TABLE 2
MEASUREMENTS OF FEM FOR KOSARAK (MINSUP=0.07%)

Thres. K_i	Mining time (second)	# frequent patterns by *MineFPTree* (P_i)	Ratio R_i
0	3341	2776266097	N/A
32	2939	1316339679	2.1
64	2146	206479285	6.4
96	1664	26795140	7.7
128	1206	2413815	11.1
160	1005	407051	5.9
192	934	86575	4.7
224	**871**	**63876**	**1.4**
256	870	58304	1.1

Let $\{K_0, K_{1,...}K_N\}$ be the set of all values of *K* where $K_i = K_{i-1}+32$; P_i is the number of frequent patterns generated by the *FP-tree mining* task when K_i is applied; and R_i is the ratio indicating the difference between P_i and P_{i-1}. R_i is computed as $R_i = P_{i-1}/P_i$,$(i = 1…N)$. According to the above observation and our extensive tests on many datasets from a well-known repository [28], the best K_i is the one satisfying the condition $R_i < 2 \ni (\nexists R_j \geq 2, \forall j > i)$. In other words, FEM will perform best at the smallest K_i where increasing *K* does not result in a sharp drop in the number of frequent patterns found by the *FP-tree mining* task. In Table 2, this condition is satisfied for K_i=224.

4.2 Automatically computing K in DFEM

It is challenging to apply the above approach to automatically find the best *K* because this value of *K* can only be specified when the mining process completes and all P_i and R_i are computed. Hence, a practical method to predict a near-optimal value of K_i based on all P_i's estimated dynamically at runtime is developed in DFEM as described in the *UpdateK* algorithm in Figure 6.

UpdateK algorithm

Input: *NewPatterns* and *Size*
Output: updated value of threshold *K*
1: **If** UpdateK is called for the first time then
2: { Create an array *P* with *N* elements
3: Initialize all P_i to zero }
4: **For** $i = 0$ to $N - 1$
5: **If** *Size* > *i*Step* **then** $P_i = P_i$ + *NewPatterns*
6: **Else** Exit Loop
7: $K = 0$
8: **For** $i = N$-1 to 1
9: **If** $R_i \geq 2$ **then** $K = (i+1)$**Step* and Exit Loop

Fig. 6 The UpdateK algorithm

In this algorithm, an array *P* with *N* elements is created and updated for every conditional pattern base processed in *MineTree* (Section 3.3) where *N* is the number of K_i being considered. Then, the best *K* is computed by finding the smallest K_i satisfying the condition in Section 4.1. We choose default values of *N*=9 and *Step*=32 where *Step* is the distance between K_i and K_{i-1} so that the best *K* is in the range 0-256 which is found the best range in our studies and allows only small conditional pattern base is transformed to in TID bit vectors (Section 3.4). For *Step*=32, the maximum size of TID bit vectors limited by *K* is a multiple of 32 (4 bytes) has better memory utilization. This distance also guarantees that applying the condition in Section 4.1 helps to specify a near-optimal of K. In Figure 6, *NewPatterns* indicates the number of new frequent patterns and is equal to the number of items in conditional pattern base C; *Size* is the size of *C*.

5 Experiments and performance study

We benchmark DFEM, FEM and five popular frequent pattern mining algorithms. DFEM is then studied in-depth to show how well it adapts to data characteristics.

5.1 Experimental setup

Software: Seven algorithms in our experiments includes DFEM, Apriori [1], Elcat [10], FP-growth [11], FP-growth* [18], AIM2 [21] and FEM [27]. DFEM and FEM are implemented using the optimizing techniques introduced in [27]. The state-of-the-art implementations of other five algorithms are obtained from [28] and [29].

Hardware: The seven algorithms tested on a machine with dual AMD Opteron 2427 processors, 2.2GHz, 24GB memory and 160 GB hard drive running CentOS 5.3, a Linux-based distribution. We used g++ for compilation.

Datasets: Five real datasets with various characteristics (Table 3) used in our benchmark were obtained from the Frequent Itemset Mining Implementations Repository [28], a well-known repository for frequent pattern mining.

TABLE 3
DATASETS AND THEIR PROPERTIES

Datasets	Type	# Items	Average Length	# Transactions	Size (MB)
Chess	Dense	76	37	3196	0.31
Mushroom	Dense	119	23	8124	0.56
Accidents	Moderate	468	33.8	340183	33.8
Retail	Sparse	16470	10.3	88126	3.79
Kosarak	Sparse	41271	8.1	990002	32.3

5.2 Performance comparison

The experimental results (Figure 7) show that DFEM runs stably and outperforms the popular algorithms for all tested sparse and dense dataset, while the other algorithms behave differently for different datasets. DFEM also performs better than FEM, our previous developed algorithm because DFEM adapts better to data characteristics. Apriori runs slowest on three datasets Chess, Accidents, Kosarak but it does better than FP-growth* for Mushroom as well as Eclat, FP-growth* and AIM2 for Retails dataset. Eclat works better than the others except AIM2, FEM and DFEM on the dense datasets. However, for the sparse datasets such as Retail and Kosarak, Eclat runs slower than most of the others. Compared to Eclat, two algorithms FP-growth, FP-growth* run faster for the dense datasets but slower for the sparse ones. AIM2, a variant of Eclat, performs well for some sparse and dense datasets but worse for the other ones. This experiment demonstrates the time efficiency of DFEM for both sparse and dense datasets.

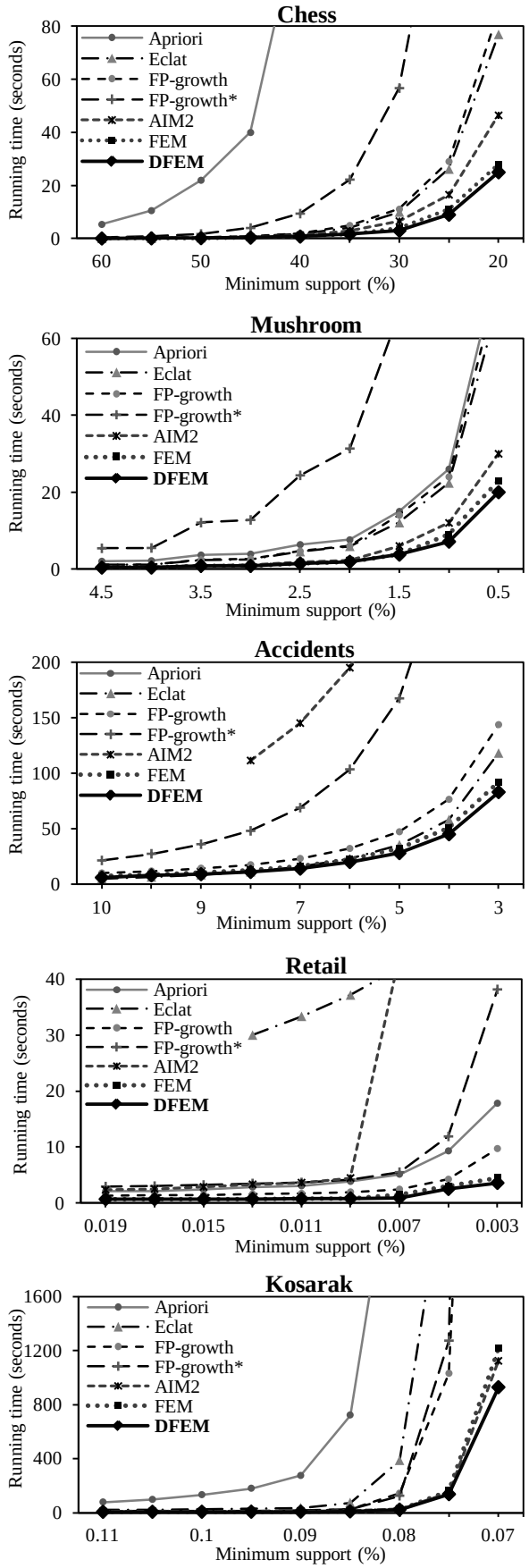


Fig. 7 Performance comparison of DFEM and other algorithms

5.3 Analyzing performance merits of DFEM

To provide insight into the performance merits of DFEM, the mining time of the *FP-Tree mining* task and the *TID bit vector mining* task were measured separately to observe the time distribution of each mining task in total mining time. The experimental results on five selected datasets are reported in Table 4. We found that DFEM distributes mining workload to its two mining tasks dynamically depending on data specific. From these results, *TID bit vector mining* is responsible for over 90% of the mining time for the dense datasets like Chess and Mushroom because the shape of the FP-tree of dense datasets is usually compact and most conditional pattern bases satisfy the condition to switch from *FP-Tree mining* to *TID bit vector mining*. In contrast, for the sparse datasets like Retail and Kosarak, *FP-Tree mining* is responsible for 90% - 99% of the mining time because many large FP-trees are generated and most of them do not satisfy the switching condition. For the Accidents dataset whose density is moderate, the mining time distribution of DFEM is balanced between the two mining tasks. It must be noted that the mining time distribution does not indicate the amount of work. In fact, the *TID bit vector mining* task using faster bitwise operations and more cache-friendly data layout will process larger amounts of data than *FP-Tree mining* does in a same unit of time.

TABLE 4
MINING TIME DISTRIBUTION BETWEEN TWO MINING TASKS OF DFEM

Datasets	Type	Minimum support (%)	FP-Tree mining time (%)	TID vit vector mining time (%)
Chess	Dense	40	4	96
Mushroom	Dense	2.5	3.5	96.5
Accidents	Moderate	5	55	45
Retail	Sparse	0.011	92	8
Kosarak	Sparse	0.09	95.5	4.5

In addition, the mining time distribution changes when the minimum support varies (Figure 8). As the minimum support is set to lower levels, more small conditional FP-trees are generated and hold the condition to switch from FP-tree mining to *TID bit vector mining* which makes the mining time percentage of *TID bit vector mining* increases as the minimum support is reduced.

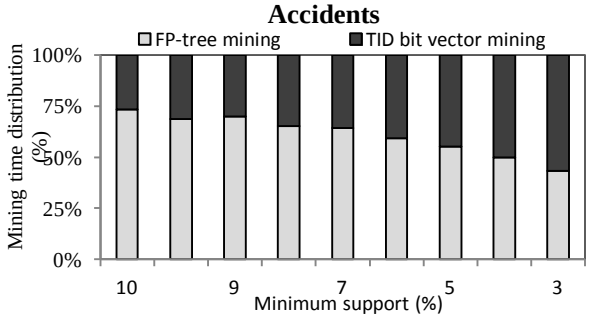


Fig. 8 Mining time distribution of DFEM for Accidents dataset

In conclusion, DFEM have the ability to switch between strategies at runtime by distributing the mining

workload to the appropriate strategy that best fit the data characteristics.

6 Conclusion and future work

In this paper, we present DFEM, a new algorithm for frequent pattern mining that can adapt its mining behavior to data characteristics to efficiently find all frequent patterns from both sparse and dense databases. Compared to FEM, our previously developed algorithm for this mining task, DFEM can automatically specify a good runtime threshold K used to switch between the two mining tasks. The experimental results show that DFEM significantly improve the performance of mining frequent patterns and outperforms other well-known algorithms for different database types. In future work, we will study parallel approaches for implementing FEM and DFEM on parallel and distributed systems because memory limitation and computational time are the major obstacles to deploying any sequential frequent pattern mining algorithm on very large scale databases.

7 References

[1] R. Agrawal, R. Srikant, "Fast Algorithms for Mining Association Rules," *Proc. of the 20th Int. Conf. on Very Large Databases*, pp. 487-499, 1994.

[2] S. Brin, R. Motwani, C. Silverstein, "Beyond Market Basket: Generalizing Association Rules to Correlations," *Proc. ACM SIGMOD Management of Data*, vol. 26, issue 2, pp. 265-276, Jun. 1997.

[3] C. Silverstein, S. Brin, R. Motwani, J. Ullman, "Scalable Techniques for Mining Causal Structures," *J. Data Mining and Knowledge Discovery*, vol. 4, issue 2-3, pp. 163–192, July 2000.

[4] R. Agrawal, R. Srikant, "Mining Sequential Patterns," *Proc. Data Engineering*, pp. 3–14, 1995.

[5] H. Mannila, H. Toivonen, and A. I. Verkamo, "Discovery of Frequent Episodes in Event Sequences," *J. Data Mining and Knowledge Discovery*, vol. 1, issue 3, pp. 259-289, Sep. 1997.

[6] J. Han, G. Dong, Y. Yin, "Efficient Mining of Partial Periodic Patterns in Time Series Database," *Proc. IEEE Data Engineering*, pp. 106-115, Mar. 1999, doi: 10.1109/ICDE.1999.754913.

[7] J. Han, H. Cheng, D. Xin, X. Yan, "Frequent Pattern Mining: Current Status and Future Directions," *J. Data Mining and Knowledge Discovery*, vol. 15, issue 1, pp. 55-86, Aug. 2007.

[8] D. Burdick, M. Calimlim, J. Flannick, J. Gehrke, T. Yiu, "MAFIA: A Maximal Frequent Itemset Algorithm," *IEEE Trans. on Knowledge and Data Engineering*, vol. 17, no. 11, pp. 1490–1504, Nov. 2005, doi: 10.1109/TKDE.2005.183

[9] H. Li, Y. Wang, D. Zhang, M. Zhang, E. Chang, "PFP: Parallel FP-Growth for Query Recommendation," *Proc. 2008 ACM Recommender systems*, pp. 107-114, 2008.

[10] M. Zaki, S. Parthasarathy, M. Ogihara, W. Li, "New algorithms for fast discovery of association rules," *Proc. Knowledge Discovery and Data Mining*, pp. 283-286, 1997.

[11] J. Han, J. Pei, Y. Yin, "Mining Frequent Patterns without Candidate Generation," *Proc. Management of Data*, vol. 29, issue 2, pp. 1-12, Jun. 2000.

[12] JS. Park, MS. Chen, P. Yu, "An Effective Hash-based Algorithm for Mining Association Rules," *Proc. ACM SIGMOD Management of Data*, vol. 24, issue 2, pp. 175–186, May 1995.

[13] H. Toivonen, "Sampling Large Databases for Association Rules," *Proc. Very Large Databases*, pp. 134–145, 1996.

[14] S. Brin, R. Motwani, JD. Ullman, S. Tsur, "Dynamic Itemset Counting and Implication Rules for Market Basket Analysis," *Proc. ACM SIGMOD Management of Data*, vol. 26, issue 2, pp. 255–264, 1997.

[15] A. Fiat, S. Shporer, "AIM: Another Itemset Miner," *Proc. Frequent Itemset Mining Implementations*, 2003.

[16] M. J. Zaki, K. Gouda, "Fast Vertical Mining Using Diffsets," *Proc. ACM SIGKDD Knowledge Discovery and Data Mining*, pp. 326-335, 2003.

[17] C. Borgelt, "An Implementation of the FP-growth Algorithm," *Proc. OSDM Frequent Pattern Mining Implementations*, Aug. 2005.

[18] G. Grahne, J. Zhu, "Efficiently Using Prefix-trees in Mining Frequent Itemsets," *Proc. Frequent Pattern Mining Implementations*, pp 123–132, 2003.

[19] J. Pei, J. Han, H. Lu, S. Nishio, S. Tang, D. Yang, "Hmine : Hyper-structure Mining of Frequent Patterns in Large Databases," *Proc. IEEE Data Mining*, pp. 441–448, Nov. 2001, doi: 10.1109/ICDM.2001.989550

[20] B. Racz. "nonordfp: An FP-growth Variation Without Rebuilding the FP-tree," *Proc. IEEE ICDM Workshop on Frequent Itemset Mining Implementations*, Nov. 2004.

[21] S. Shporer, "AIM2: Improved Implementation of AIM," *Proc. IEEE Frequent Itemset Mining Implementations*, Nov. 2004.

[22] L. Schmidt-Thieme., "Algorithmic Features of Eclat," *Proc. IEEE Frequent Itemset Mining Implementations*, Nov. 2004.

[23] L. Liu, E. Li , Y. Zhang, Z. Tang, "Optimization of Frequent Itemset Mining on Multiple-core Processor," *Proc. of the 33rd Int. Conf. on Very Large Databases*, pp. 1275-1285, 2007.

[24] W. Li, A. Mozes, "Computing Frequent Itemsets Inside Oracle 10g," *Proc. of the 30th Int. conf. on Very Large Databases*, pp. 1253–1256, 2004.

[25] C. Utley, "Introduction to SQL Server 2005 Data Mining," *Microsoft SQL Server 9.0 technical articles*, available at: http://technet.microsoft.com/en-us/library/ms345131.aspx, Jun. 2005.

[26] T. Yoshizawa, I. Pramudiono, M. Kitsuregawa, "SQL Based Association Rule Mining Using Commercial RDBMS (IBM db2 UBD EEE)," *Proc. Data Warehousing and Knowledge Discovery*, pp. 301–306, 2000.

[27] L. Vu, G. Alaghband, "A Fast Algorithm Combining FP-Tree and TID-List for Frequent Pattern Mining," *Proc. IEEE of the 2011 Inf. Conf. on Information and Knowledge Engineering*, pp. 472-477, Jul. 2011, Las Vegas, NV, USA.

[28] "Frequent Itemset Mining Implementations Repository," *Workshop on Frequent Itemset Mining Implementation*, 2003-2004, available at http://fimi.ua.ac.be

[29] Christian Borgelt, "Frequent Pattern Mining Implementations," available: http://www.borgelt.net.

[30] B. Goethals, M. J. Zaki, "Advances in Frequent Itemset Mining Implementations: Report on FIMI'03," *ACM SIGKDD Explorations Newsletter - Special issue on learning from imbalanced datasets*, Vol. 6 Issue 1, pp. 109-117, June 2004, New York, NY, USA.

Organizing Linked Data Quality Related Methods

Philippe A. MARTIN
ESIROI I.T., EA2525 LIM, Uni. of La Réunion, Sainte Clotilde, France
(+ adjunct researcher of the School of ICT, Griffith Uni., Australia)

Abstract - *This article presents the top-level of an ontology categorizing and generalizing best practices and quality criteria or measures for Linked Data. It permits to compare these techniques and have a synthetic organized view of what can or should be done for knowledge sharing purposes. This ontology is part of a general knowledge base that can be accessed and complemented by any Web user. Thus, it can be seen as a cooperatively built library for the above cited elements. Since they permit to evaluate information objects and create better ones, these elements also permit knowledge-based tools and techniques – as well as knowledge providers – to be evaluated and categorized based on their input/output information objects. One top-level distinction permitting to organize this ontology is the one between content, medium and containers of descriptions. Various structural, ontological, syntactical and lexical distinctions are then used.*

Keywords: Knowledge quality evaluation, Knowledge sharing ontology, Knowledge organization and best practices

1 Introduction

How should data or knowledge be represented and published so it can most easily be retrieved, re-used and managed? Then, how to compare or evaluate knowledge statements, knowledge bases (KBs), knowledge management techniques, KB management systems (KBMSs) and knowledge providers? Many complementary or alternative "knowledge sharing supporting *elements*" – and kinds of elements – have been proposed to provide partial answers to these research questions. For example: Semantic Web approaches [1-2], knowledge sharing languages (e.g., those provided by the W3C and more general ones [3-6]), ontologies [7-13], methodologies [14-15], best practices or design patterns [16-18], categories of evaluation criteria or measures [19-22], knowledge quality evaluation queries [23-25], benchmarks [26-27], techniques [28-29], software, etc. (these references are given to illustrate each "kind of element"; many of them will also later be referred to in this article; giving them *now* permits to group them by "kind" in the "references" section of this article).

However, it is difficult for a knowledge provider or a KBMS developer to know about all these elements or their sub-elements, to compare them, choose between them, combine them and have a *synthetic organized view* of what can or should be done for knowledge sharing purposes. Indeed, these elements often do not use similar terminologies or categorizations, and *no ontology or library* has been proposed *to compare, index, organize and generalize* these elements (which are at various levels of abstraction and may be contradictory). *The top-level ontology presented in this article is a step in that direction. Thus, the goal of this article is to show how the problem cited at the beginning of this paragraph can be addressed.* This article refers to this ontology or library as "this knowledge criteria/quality ontology" or simply "this (top-level) ontology". It is currently focused on elements related to Linked Data [2].

As with any other ontology, the bigger *and more organized* it will become, the more useful it will be for the above cited knowledge sharing and retrieval tasks. (For the knowledge operationalization tasks, bigger is no longer better but eases the selection of knowledge for modules of relevant sizes and content for an application; this article and its top-level ontology only address knowledge sharing and retrieval tasks). This ontology is part of a KB published on-line via the WebKB knowledge server (usable at http://www.webkb.org). This KB – and hence the ontology presented in this article – can be cooperatively extended by any Web user via this server. To enable this, WebKB uses an abstract model and editing protocols [29] allowing its KB to be *consistent* and organized even though knowledge statements come from different sources and hence can *contradict* each other. Because of space restrictions this model is only quickly introduced in this article (it is not described) but this knowledge quality ontology exploits it and categorizes some of its features.

This ontology is only about information objects. Indeed, once information objects can be evaluated, knowledge management tools and techniques can be compared or evaluated with respect to (w.r.t.) the qualities of the information objects that they allow as input and output, or lead their users to produce. Similarly, knowledge providers can be evaluated w.r.t. the information objects they have provided. Thus, unlike the "Semantic Web (SW) Topics Ontology" [10], this ontology does not attempt to semantically organize – i.e., does not attempt to use specialization relations, "part of" relations or other relations to organize – knowledge management tools, techniques or processes (e.g., tools and techniques for knowledge extraction, retrieval, matching, merging, representation, inferencing, validation, edition, annotation, modularization and publishing). This would be a huge task and, as for example illustrated by the "SW Topics Ontology", these processes are so intertwined that they are difficult to distinguish and organize in a *scalable* way, i.e., in a systematic and non-arbitrary way within a specialization hierarchy and a part-of hierarchy. Here, "non-arbitrary" implies the use of conceptual distinctions – and especially, partitions – that are clear enough to lead different persons to categorize a same thing at a same place in a specialization/part-of hierarchy (note: a hierarchy does not have to be a tree). Such distinctions and hierarchies significantly reduce implicit redundancies [14].

The uppermost conceptual distinction used by this ontology to permit non-arbitrary categorizations of information objects is the clear partition of information objects into either description-content, description-medium or description-container objects. In this article, *description-content* objects are conceptual categories, as well as formal/informal terms or statements referring to or defining these categories. They are interpretations or abstractions of a (real or imaginary) situation or object. E.g.: abstract models, ontologies, terminologies, languages and any of their sub-elements (e.g., the concept/relation types of RDF and OWL). D*escription-medium* objects are concrete model objects permitting to visually/orally/... present description-content objects. E.g.: graphical interface objects and syntax/style objects such as those specified by XML, CSS and XSLT. *Description-container* objects are the other information objects, i.e., non-physical objects permitting to store and manage description-content and description-medium objects. E.g.: files, file repositories, distributed databases and file servers.

The sections 2, 3 and 4 are respectively about the evaluation of description-content, description-medium and description-container objects. These sections relate, organize and generalize knowledge sharing best practices and quality criteria/measures from various sources. Some categories from each of the above referred articles are included in this quality ontology. [17-22] include the most complete lists of high-level categories that seemed to exist so far for Linked Data. All their categories are integrated in the quality ontology. The top-level of this ontology may be seen as validated by the fact it correctly organizes and generalizes all the quality-related categories that the author has found so far (this is a kind of "validation by usage"). On the other hand, this was obvious given the way this ontology was designed. The difficult part was to find this design. This ontology may also be seen as validated by the fact it follows the strongest best practices it categorizes (the ones that include or imply the weaker or more elementary best practices, e.g., the fact that a KB should at least be consistent). In [17-22], elements are only slightly categorized. This is shown by Section 5 which gives the four most organized quality related categorizations that seemed to exist for Linked Data so far. These categorizations are essentially only two levels deep and not always intuitive. Additional conceptual distinctions would be interesting, especially if they are "non-arbitrary" (under this condition, the more categories, the better).

2 Description content quality

This article cannot present the whole "quality ontology". It can only show its top-level and its principles, i.e., the way this ontology manages to organize the *main kinds* of methodological elements, best practices, quality characteristics (e.g., evaluation criteria, quality dimensions, the "data quality indicators" of [22], ...) and quality measures (e.g., the "scoring functions" and "assessment metrics" of [22], ...) that have been proposed for knowledge sharing purposes. This article only shows important elements of a *subtype hierarchy of quality measuring functions* on information objects, with the function result being a value (typically, numerical or boolean). Indeed, all quality related categories and subtype hierarchies can be automatically *derived* from the above subtype hierarchy. E.g., *relations* can be derived from boolean functions (Figures 3 includes an example) and, from the above subtype hierarchy, it is possible to derive the one for *quality characteristics* and the one for *"statements that have a certain (kind of) quality measure"* (alias, "statements that follow a certain (kind of) best practice").

There are many ways to categorize quality evaluations, e.g., according to the kind of objects they evaluate, and whether or not they take into account certain lexical, structural or semantic best practices. Figure 1 shows one intuitive uppermost categorization. In this article, indented lists show subtype hierarchies. In all such indented lists below, the XML namespace shortcut can be used but the prefix "pm:" is left implicit. "LDpattern:" is for [18], "LD:" for [20], "SF:" for [21] and "PD:" for [22-23]. C++/Java-like comments are used. Relation identifiers use nominal expressions and follow the common "graph reading convention", i.e., the last argument is the destination of the relation. Thus, a binary relation R(X,Y) can be read "<X> has for <R> <Y>" (notes: <X> and <Y> may include quantifiers; furthermore, a relation "X has for subtype Y" may also be read "any instance of Y is also an instance of X"). For functional relations, the last argument is the function result.

Figure 2 gives subtypes to pm:description_content_quality. Two points explain those subtypes. First, one handy partition for description-content_semantic-quality functions is the distinction between those that give "correctness" values for the evaluated object and those checking that it includes certain things. Second, for each kind of evaluated source object, there are various ways to categorize i) functions that evaluate certain aspects of this kind of objects (e.g., the "correctness" and "conformity" aspects), ii) functions that evaluate "*related* objects", and iii) functions that differently aggregate the values returned by these functions. In this article, within the names of the last two kinds of functions, "description" is abbreviated by "descr" in order to make the hierarchy more readable. The adjective "*related*" refers to *actual or potential/allowed* relations. E.g., RDF (a description content) *allows* various kinds of textual or graphical notations (description media) – some being standards, some not – even if most RDF-based tools (description containers) only work with RDF/XML. Some evaluation functions may for example *better* rate an RDF-based tool that can handle *more* notations (e.g., by calling external translation tools).

```
quality  //function on an object with possibly other arguments;
         //this function returns a value (numerical, boolean, ...)
  content_based_quality  //at least based on the object content
  meta-statement_based_quality  //ie., on meta-statements on the object
    rating_based_quality  //at least based on ratings
```

Figure 1. Important top types to organize quality measuring functions

```
description-content_quality  //subtype of above function pm:quality
  correctness  //one main kind of description-content_quality;
               // Figure 3 gives some s important subtypes
  conformity  //another main kind; Figure 5 gives important subtypes
  quality_of_this_descr_content  //to evaluate the source object
                                 // on all its criteria
    descr_content_quality_of_this_descr_content  //content-related
                                                 // aggregations
    quality_of_descr_media_related_to_this_descr_content
    quality_of_descr_containers_related_to_this_descr_content
```

Figure 2. Important ways to evaluate description content quality

```
correctness   //of the evaluated object (statement or term
              // referring to a statement)
  LD:accuracy   //factual correctness of a statement (which
                // should be a belief) w.r.t. the world
  consistency   //reports all or some inconsistencies and
                // implicit contradictions
    consistency_of_this_statement_wrt_this_one (ST,ST -> boolean)
                  //this signature states that this function is boolean
                  // and has exactly 2 statements as arguments;
                  //one relation derivable from this function is:
                  // pm:statement_consistent_with_this_one (ST,ST)
      consistency_and_non-redundancy_of_this_statement_wrt_this_one
                  //signatures are inherited when there is no ambiguity
    consistency_of_this_KB (ST -> boolean)
      consistency_of_the_RDF_KB  //the KB must be an RDF KB
        consistency_of_SKOS_relations  //the measures of [24] are
                                       // subtypes of this type
          consistency_of_a_RDF_KB_tested_via_a_SPARQL_query
                                       // as in [23] and [25]
      LD:internal_consistency_fct  //SF:consistency_fct seems to
                                   // be an alias of this type
      LD:modeling_correctness_fct  //tests the correctness of the
                  // "logical structure of the data"; LD and SF do not
                  // precise if these last 2 types are dimensions or
                  // functions; this is why "_fct" is a suffix here;
                  // other relations and dimensions can be derived
    substatements_of_this_1st_statement_that_are_inconsistent_\
with_this_2nd_statement (ST,ST -> set)
    consistency_ratio  //no restriction on the arguments but the
                       // result is a number (a ratio)
      consistency_ratio_of_such_a_statement_in_this_statement
          (ST,ST -> number)  //"such a statement": the 1st argument
      consistency_ratio_of_all_substatements_in_this_statement
          (ST -> number)  //contextualized substatements
        consistency_ratio_of_this_KB (ST -> number)
      consistency_ratio_of_relations_on_this_term_in_this_statement
          (term -> number)
      consistency_ratio_of_this_relation_on_this_term_in_this_statement
          (ST,term -> number)
        PD:consistency  //"number of non-conflicting frames"
                        // divided by the "number of frames"
```

Figure 3. Important types of functions to evaluate correctness

Figure 3 gives some specializations for the first subtype of pm:description_content_quality. So far, all current quality measures related to Linked Data seemed to use the whole KB (data set) as *implicit argument*. They may thus be simpler to call or to understand but this is a loss of generality. Furthermore, most of these measures only work on "frames" ("objects" in object-oriented approaches), i.e, on the set of relations from a term. They do not work on *any kind of statement*. The hierarchy in Figure 3 shows how different but related evaluation functions and relations can be organized and generalized. Concept types can be derived too. An important one is pm:Statement_consistent_and_non-redundant_with_any_other_one_in_the_KB. Indeed, if a KBMS checks that each statement is of this type before allowing its insertion it into the KB, every statement has a place in the specialization hierarchy of this KB and can thus be automatically compared to other statements. This is enforced by the KB editing/sharing protocols of WebKB.

```
(pm%def_fct  ;;one of the operators for defining term, here a function
   pm%consistency_ratio_of_such_a_statement_in_this_statement
     (?s1 ?s2)  ;; parameters;  "?" is a KIF prefix for variable names;
                ;;as in all evaluation functions, the first argument
                ;;  is the evaluated object;  ?s2 may be a whole KB
   (div  (pm%cardinality  ;;size of the set returned by "setofall"
           (setofall ?s (and  (pm%substatement ?s2 ?s)  (=> ?s ?s1)
                              (not (=> ?s2 (not ?s))) )))
         (pm%cardinality (pm%substatements ?s2)) ))
```

Figure 4. Full definition of one of the above functions, in KIF

[23] and [25] implement some quality measures (for whole RDF-based KBs) via SPARQL queries and SPIN rules. As noted in Figure 3, some of these measures are about consistency. They permit to check some aspects that KB building tools already verify or some complementary aspects. More classic and powerful consistency checking or other quality measurement cannot be done – or would be too complex and long to be done – via SPARQL queries. They require inference engines and definitions in expressive languages such as KIF [5]. The Common Logic ISO standard and its CLIF notation [6] are adaptations of the model and syntax of KIF but without features for specifying meta-statements, definitions and (monotonic or not) inference rules. This is why the data model (alias, meta-ontology) of WebKB is defined in KIF and why some definitions of its quality related ontology are in KIF. As illustrated by Figure 4, the functions referred to in this article are relatively easy to write in KIF. This figure shows the use of "pm%def_fct", one of the primitive definition operators of the data model of WebKB, fully defined in KIF. These operators also permit to check and handle lexical/namespace and semantic contextualizations in a KB. "%" is used instead of ":" for XML-like namespace shortcuts because "%" and "#" have other meanings in KIF. "Semantic contextualization" refers to the embedding of a statement within a meta-statement for specifying information without which the statement may be false, e.g., information about the time, place and source of the content of the statement. Every statement must have at least one source (author, source document, etc.). This permits to consider "every assertion that is not a definition" as a belief. This is the main reason the KB can stay consistent while still allowing contradictory beliefs.

The KB editing/sharing protocols of WebKB also enforce the statement of relations – such a *pm:**corrective**_refinement* – between contradictory beliefs. This avoids implicit partial/total redundancies and permits people or applications to choose between these contradictory beliefs when choices have to be made. Thus, no arbitrary selections has to be made *a priori* by the KB owners. E.g., a user may specify that "when browsing the KB or within the results of his knowledge extraction queries" he wants to see only the most specialized *corrections* from certain kinds of sources. Furthermore, WebKB allows "default beliefs" and provides some default beliefs. E.g., a default belief for any user in WebKB is that, among contradictory statements, he believes the ones that *correct* the others without being contradictory between themselves. Since quality related measures can be specified as definitions or as beliefs, by creating or overriding some defaults beliefs, a user can easily specify which measures he believes in and wants to use, and how they should be combined. For more details on semantic/lexical contextualizations, default beliefs, and how to combine them, see the KIF formalization of the WebKB data model at http://www.webkb.org/kb/it/KSmodel.html

conformity //reports on the existence/number of certain things or //patterns (thus, even SF:amount_of_data is a subtype of it)
- ***conformity_of_this_statement_wrt_this_requirement*** (ST,ST -> boolean)
- ***ratio_of_conformity_to_this_requirement_in_this_statement*** (ST,ST -> number)
- ***ratio_of_conformity_of_the_KB*** //no argument restriction here
  - LD:modeling_granularity (-> number) //no argument
  - PD:structuredness //e.g., PD:coverage (number of objects // with all relations of a schema) and // PD:coherence (average of coverage for all terms)
  - PD:completeness //alias, LD:completeness (do all required // terms/relations exist?)
    - PD:intensional_completeness //ratio (percentage) of // required relations in the KB
    - PD:extensional completeness //ratio of required terms
    - PD:LDS_Completeness //ratio of terms with a given relation
  - PD:relevancy //alias LD:Boundedness, ratio of data relevant // for an application
  - PD:verifiability //existence of information to check for //correctness; examples of subtypes: PD:traceability, // PD:provability, PD:accountability; the following two // best practices are related to these subtypes: // - "providing another KB for tools that cannot perform // complex inferences" (LDpattern:materializing_inferences) // - "transforming the KB to conform to some models" // (LDpattern:transformation_query)
- SF:validity //no syntax errors, …; PD:verifiability is very // related; PD:validity is a subtype
- ***representation_quality***
  - ***organization*** //for formal and informal objects in the KB //WebKB permits to organize both
    - ***at_least_minimal_organization*** //as defined and enforced in //WebKB; many other subtypes can be defined
    - ***reachability*** //PD:reachability when the object is a KB
      - out-relations //from the object; for a whole KB: // PD:external_links, PD:outdegree, … //the more out-relations, the better: this is the // 4th basic rule for Linked Data [16]; the more widely // known/deployed the target objects, the better
      - in-relations //to the object; for a KB: PD:indegree, …
    - ***non-redundancy*** //e.g., PD:intensional_conciseness, …
  - ***expressiveness_economy*** //avoidance of expressive constructs // when this does not bias knowledge representation and // reduce knowledge matching/inferencing/readability
  - ***modeling_uniformity*** //e.g., checks some lexical,structural // or ontological conventions
    - LD:directionality //checks the consistency in the // direction of relations
    - ***use_of_the_graph-oriented_reading_convention*** //as for the // 5 above types, it is important for readability
    - ***conformity_to_an_abstract_model_or_ontology_or_methodology***
      - conform_to_Ontoclean //checks that the object (or each of // its sub-objects) is instance of at least one of the Ontoclean // 2nd-order types: (semi/anti/totally_)rigid_thing, etc.
      - use_of_a_standard_model //3rd basic rule for Linked Data // but for abstract models only
  - ***quality_of_the_representation_of_terms*** //see Figure 6

Figure 5. Important subtypes of functions to evaluate conformity

Figure 5 – and, its continuation, Figure 6 – show subtype relations between types of functions checking that within an object certain elements exist and are conform to a certain pattern. The various subtypes are semantically close. The first listed subtype can be re-used to write the other ones. This subtype hierarchy shows that the "current categorizations for Linked Data quality criteria and measures" (LD, PD, SF, ...) only cover particular cases. Thus, the current implementations of (some of) these measures also only cover particular cases. To save space, there is no repetition of types in this hierarchy (this applies to the next hierarchies too). Some of the types could clearly also appear at other places. The comments give some explanations for each of the types. The ones in bold and/or italics are the most important for categorization or re-use purposes.

- ***quality_of_the_representation_of_terms*** //as in end of Figure 5
  - identification_by_properly_formed_URIs //checks that // objects are identified by HTTP URIs that can be // dereferenced by agents to find further information; // these are first two rules of Linked Data [16]; // [18] gives some specializations to these best practices
  - ***following_of_naming_conventions*** //use of nouns, of a // loss-less naming style, …
  - LD:referential_correspondence //consistency and // non-redundancy of identifiers
  - LD:typing //checks that nodes are first-order typed // entities, not just strings, hence checks the // "Link Not Label" best practice [18]
  - PD:vocabulary_understandability //checks that terms have // human readable labels, …
    - LD:intelligibility //alias SF:comprehensibility? These two // types appear to be only about terms
    - PD:internationalization_understandability //checks that // the language is specified
- ***quality_of_existing_or_derivable_relations***
  - use_of_binary_relations_only //since this helps knowledge // matching and precision
  - ***quality_of_existing_or_derivable_meta-statements*** //and hence // relations from statements
    - quality_of_existing_or_derivable_contexts //temporal ones, // spatial ones, modal ones, …
      - provenance //checks that the sources (agents/files) are // represented (LD:Attribution) and the creation dates // too (LD:History); LD:Authoritative is for checking // if the author is a credible authority in the domain
      - ***loss-less_integration*** //checks that the semantics of // source objects was not changed; the data model of // WebKB and its protocols permits such an integration
      - PD:timeliness //alias SF:timeliness and LD:Currency; is // the object is up-to-date? E.g., PD:newness (timely // creation?) and PD:freshness (timely update?)
      - SF:licensing //alias, LD:licensed; to check for an open // license, use PD:openness
      - security //checks for signatures, encryption, // maintainability (LD:sustainable), …

Figure 6. Important functions to evaluate the quality of the representation of terms

3 Description medium quality

Description-medium quality functions evaluate the textual/graphic/... presentation that some (kinds of) description-medium objects permit for some (kinds of) description content objects in some (kinds of) description containers. The more the "presentation is distinguished from content" and the more structured (fine-grained) the content, the more the presentation can be finely adapted for different kinds of users and by the end-users. To that end, the W3C advocates the use of XML-based languages (e.g., RDF/XML) as well as CSS, XSLT and GRDDL. This last language indicates which XSLT scripts can be used to translate some knowledge published in some XML-based KRLs into other knowledge representation languages (KRLs). Fresnel [28] may be seen as a kind of advanced CSS for RDF-based KRLs.

Presentation evaluation functions may for example give high values to graphical/textual/audio/... interfaces composed of fine-grained objects with "rich contextual menus". For each object, such menus would list i) presentation attributes/commands for this object, and ii) semantic relations/commands from/to/about this object to ease navigation, querying or updates. Although a syntax is clearly a presentation object, neither XML nor any current KRL seem to allow people to adapt their syntaxes, e.g., via the setting of some variables or the use of a notation ontology. Yet, this approach would be more flexible and easier to use than GRDDL, and hence can be used as one criteria by description-medium quality evaluating functions. Figure 7 gives a top-level specialization hierarchy for such functions. The general comments on the previous hierarchies also apply here: conventions, abbreviations, rationale for the specialization relations, etc.

```
description-medium_quality  //subtype of function type pm:quality
  quality_of_this_descr_medium  //to evaluate the source
        // object on all its criteria (→ aggregations of measures)
    descr_medium_quality_of_this_descr_medium
    quality_of_the_descr_content_related_to_this_descr_medium
    quality_of_the_descr_containers_related_to_descr_medium
  use_of_standard_formats  //for used KRLs (RDF/XML, …),
       // for character encodings, graphics (SVG, …), … (see w3.org);
       //3rd basic rule of Linked Data but for concrete models only
  use_of_structured_formats  //e.g., an HTML presentation with or
                             // without RDFa statements
    use_of_formats_distinguishing_structure_from_presentation
        //e.g., XML but note that XML does not permit its users
        // to adapt its notation via the setting of some values
    use_of_notations_that_can_be_adapted_by_the_user
        // unlike XML and almost all notations
    use_of_machine-understandable-formats
    use_of_formats_that_have_an_interpretation_in_some_logic
      PD:format_interpretability  //aggregation of measures on
                          // qualities of formats proposed by a KB
        PD:human_and_machine_interpretability  //e.g., N3 can
                          // be more easily read than RDF/XML
      format_structural_quality  //subtypes on the next column
  format_concision  //e.g., N3 is more concise than RDF/XML
  format_uniformity  //reports on the extent to which similar
                // things can be (re)presented in similar ways (from
                // a software viewpoint and/or from a person viewpoint)
    SF:Uniformity  //pm:format_uniformity for a whole KB
  performance_of_this_format_for_this_task
    (description_medium, task -> value)  //function signature
```

```
/* Figure 7 continues here to detail the following subtype branch */
      format_structural_quality  //see the previous column
        format_abstract-expressiveness  //the expressiveness of
          // its abstract model (→ first-order_logic, …, kinds of
          // possible quantification (note: KIF allows to define
          // all kinds of relations to represent numerical quantifiers
          // but has no predefined keywords for them; thus, numerical
          // quantifiers defined by different users will be hard to match
          // (especially via simple graph-matching based techniques);
          // hence, KIF is expressive but low-level
        syntactic_expressiveness  //the higher the numeric result of
          // this function, the higher-level the notation can be considered
          // (for the selected criteria), i.e., the more flexible and
          // readable the format is, the more normalized/uniform
          // the descriptions are, and hence the easier to compare
          // via graph-matching these descriptions are
          syntactic_constructs_for_logical_constructs  //e.g., does
            // the format include keywords for numerical quantifiers
            // (e.g., "58%", "2 to 6") and for which kinds of them
          syntactic_constructs_for_creating_shortcuts  //kinds
            // of lambda-abstractions, …
          syntactic_constructs_for_ontological_primitives
            //e.g., for type partitions and/primitives such as those in
            // Ontoclean and extensions of them. They are needed
            // for knowledge engineering [3]. RDF is low-level: it
            // has no keywords for them but can import a
            // language ontology which has them
          referable_first-order-entities  //e.g., what can be a
            // 1st-order entity, i.e., what can be referred to via a
            // variable in the notation: concept nodes, relation nodes,
            // quantifiers, …; the more things can be 1st-order entities
            // (and hence that can be related to other things,
            // annotated, selected via a mouse, …), the better, and
            // the more formally related, the better for structuring or
            // annotation flexibility purposes;
            // from that viewpoint, an interface or notation for a KB
            // may be better than one for a database or a
            // structured document (which is then also better than an
            // unstructured one)
```

Figure 7. Important functions to evaluate the quality of a description medium

4 Description container quality

Description-containers quality functions evaluate the way a given description container – e.g., a static file or a distributed KB server – i) modularizes, stores, makes retrievable and accessible (i.e., how it "publishes") description content objects, and ii) checks or allows updates or queries on these objects. Compared to the independent and direct use of static files (e.g., RDF files), the use of knowledge servers eases knowledge modeling and reduces the implicit inconsistencies and redundancies between their knowledge statements. A KB server can also use static input/output files and offers much more flexibility than static files. It can also provide more services than those of a description-container (e.g., it can forward queries). This can be taken into account for evaluating its quality. Figure 8 gives a specialization hierarchy for description-container quality evaluating functions. The general comments on the previous hierarchies still apply.

```
description-container_quality  //subtype of function pm:quality
  quality_of_this_descr_container
    descr_container_quality_of_this_descr_container
    quality_of_the_descr_content_related_to_this_descr_container
    quality_of_the_descr_media_related_to_descr_container
    quality_of_the_processes_supported_by_this_descr_container
  storage_related_quality
    maximal_size_of_the_KB
    container_based_modularization
      static_container_based_modularization  //static file based
      dynamic_container_based_modularization  //forwarding or
                  // replication of knowledge/queries amongst KBs
      LD:connectedness  //do combined datasets join correctly?
  assertion_related_quality  //what can be added or updated, by
                                  // whom, in which language, …
    ontological_flexibility  //is the ontology fixed, i.e., is the
                                 // KB actually just a database?
      LDpattern:annotation  //are third-party resources accepted?
      LDpattern:progressive_enrichment  //ways data (model) can
                                                //be improved over time
    checking_possibilities  //what kinds of inconsistencies or
                  // redundancies or redundancies can be detected?
                  //does the server advocate best practices to its users?
  information_retrieval_related_quality  //on the whole KB or on
                                                // some of its statements
    published_or_given_metadata  //on the KB or a part of it, e.g.,
        //via a "topic" (Ldpattern:Document_Type), via the use of a
        // semantic sitemap [11], voiD or DCAT, via metadata given
        // for any object (if a user requests it) but calculated in a
        // predefined way (as with "Concise Bounded Descriptions")
        // [2], or via metadata accessible via powerful queries
    object_accessibility
      PD:accessibility  //access methods, e.g., via SPARQL, an API,
                                //                  a file (HTML,RDF)
      PD:availability  //percentage of time a given service is "up"
      SF:performance  //low latency, high throughput, only minor
                                // "performance variations", …
        PD:response_time  //e.g., for static access, SPARQL access
        PD:robustness  //average of performance over time; helped
                  // data cache and the use of LDpattern:parallel_loading
      querying_possibilities  //what can be queried, with which
                  // input/ouput languages, what privacy techniques are
                  // used, are the results ranked, filtered and merged, …
  interface_personalization  //to which extent can the input/output
     // presentation  be adapted by end-users and can take into account
     // their constraints: language, disabilities,
     //                  access from various devices (mobile ones, …),
     //                  access from various software (browsers, …), …
```

Figure 8. Important ways to evaluate a description medium structural quality

5 Some other categorizations

In order to show how this knowledge criteria/quality ontology extends, generalizes and organizes the elements of its sources, Figure 9 lists show the structure of the four most organized sources that so far seemed to exist for Linked Data, even though they are essentially only two level deep. "SF:" is for [21], "Kahn" is for [19], "LDpattern:" is for [18] and "OPD:" is for [17] (this last source has three 3-level deep categories and one 4-level deep category). The first two sources are about quality criteria, the last two are about best practices. Their categories – the ones shown below – *seem to be* concept types. To permit a maximal integration of the various sources, they have been integrated into this quality ontology via function types, as illustrated by the previous indented lists. From these functions hierarchies, the concept types hierarchies can be generated. In the following lists of Figure 9, the lowermost categories are given within comments *and* without prefix for their source. The lowermost "OPD" subtypes have several instances in the OPD library.

```
SF:Quality_criterion  //this categorization often (but not always) follows
        // the distinction between description content/medium/container
    SF:Content  //Consistency, Timeliness, Verifiability
    SF:Representation  //Uniformity, Versatility, Comprehensibility;
                //mixes criteria on descr. medium and descr. container
    SF:Usage  //Validity of documents, Amount of Data, Licencing;
                // mixes criteria on descr. content and descr. container
    SF:System  //Accessibility, Performance

Kahn:Quality_dimension  //claimed to be "the 15 most important
                                //      ones  from consumer perspective"
    Kahn:Intrinsic  //Believability, Accuracy, Objectivity, Reputation
    Kahn:Contextual  //Value-added, Relevancy, Timeliness,
                        // Completeness, Appropriate amount
    Kahn:Representational  //Interpretability, Ease of understanding,
                                // Consistency, Concision
    Kahn:Accessibility  //Accessibility,  Access security

LDpattern:Linked_Data_pattern
    LDpattern:Identifier_pattern  //Hierarchical URIs, Literal Keys,
                        // Natural Keys, Patterned URIs, Proxy URIs,
                        // Shared Keys , URL Slug
    LDpattern:Modelling_pattern  //Custom Datatype, Index Resources,
            // Label Everything, Link Not Label, Multi-Lingual Literal,
            // N-Ary Relation , Ordered List, Ordering Relation,
            // Preferred Label , Qualified Relation, Reified Statement,
            // Repeated Property, Topic Relation, Typed Literal
    LDpattern:Publishing_pattern  //Annotation, Autodiscovery,
            // Document Type, Edit Trail, Embedded Metadata,
            // Equivalence Links, Link Base, Materialize Inferences,
            // Named Graphs, Primary Topic,  Autodiscovery,
            // Progressive Enrichment, SeeAlso
    LDpattern:Application_pattern  //Assertion Query, Blackboard,
          // Bounded Description, Composite Descriptions,
          // Follow Your Nose, Missing Isn't Broken, Parallel Loading,
          // Parallel Retrieval, Resource Caching, Schema Annotation,
          // Smushing, Transformation Query

ODP:Ontology_Design_Pattern
    ODP:Structural ODP  //Architectural ODP,  ODP:Logical  ODP
      ODP:Logical ODP  //Logical_macro_ODP, Transformation_ODP
    ODP:Correspondence  ODP
      ODP:Alignment ODP
      ODP:Re-engineering ODP
        ODP:Schema_reengineering_ ODP  //Refactoring_ ODP
    ODP:Content ODP,  ODP:Reasoning ODP,  ODP:Lexico-syntactic ODP
    ODP:Presentation ODP  //Naming ODP, Annotation ODP
```

Figure 9. Other categorizations for some elements in the quality ontology

6 Conclusions

This article has presented the top-level of an ontology organizing knowledge sharing best practices, design patterns, evaluation criteria and evaluation measures in a systematic, non-redundant and scalable way (e.g., by being based on distinctions on information objects rather than on processes). Some other research works on this subject mainly proposed *lists* of categories with, sometimes, some implementations (e.g., via SPARQL). The integration of these categories into this quality ontology shows that the results of these works cover only particular cases, which could sometimes be easily generalized. This ontology also permits to have a more synthetic view of the *kinds* of things that could or should be evaluated or done during knowledge sharing, or proposed by knowledge engineering tools. This ontology can be extended by Web users via the server which hosts it [29]. It could then be used as an index for elements of other libraries or ontologies. To that end, the bigger it will become, the more useful it will be. The presented ontology is, in some senses, validated by the fact it includes – or can be specialized to include – any quality related measures or criteria that the author has come across and by the fact it follows the strongest best practices it categorizes. Such an ontology is clearly application-independent. No particular use case would further validate it.

Section 2 also quickly introduced the data model of WebKB which enables i) loss integration of knowledge from various sources, and ii) the use of "default beliefs/rules/measures" to allow the combination of simple evaluation functions into complex ones and the re-use of other agents' functions. The fact that knowledge on the Semantic Web is full of implicit contradictions and redundancies, very hard to evaluate, and often incorrect even with respect to the OWL primitives that it re-uses [27], may be an indication that KBMS developers and knowledge providers to the Semantic Web would benefit from such a data model and such an ontology of best practices and quality measures.

7 References

[1] A. Palma, P. Haase, Y. Wang, M. d'Aquin. "D1.3.1 propagation models and strategies". NeOn deliv. D1.3.1, 2007.

[2] T. Heath, C. Bizer. "Linked Data: evolving the Web into a global data space". Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1–136. Morgan&Claypool, 2011.

[3] G. Guizzardi, M. Lopes, F. Baião, R. Falbo. "On the importance of truly ontological representation languages". IJISMD 2010. ISSN: 1947-8186.

[4] P.F. Patel-Schneider, "A revised architecture for Semantic Web reasoning". PPSWR 2005 (Germany), LNCS 3703, 32–36.

[5] M.R. Genesereth, R.E. Fikes, "Knowledge Interchange Format". Version 3.0. Technical Report Logic-92-1, Stanford Uni., 1992.

[6] P. Hayes, C. Menzel, J. Sowa, T. Tammet, M. Altheim, H. Delugach, M. Gruninger. "Common Logic (CL): a framework for a family of logic-based languages". ISO/IEC IS 24707:2007.

[7] A. Farquhar, R. Fikes, J. Rice. "The Ontolingua server: a tool for collaborative ontology construction". International Journal of Human-Computer Studies, 1996.

[8] S. Borgo, C. Masolo. "Ontological foundations of DOLCE". Handbook on Ontologies, Springer, 361–382, 2009.

[9] N. Guarino, C. Welty. "Evaluating ontological decisions with OntoClean". Comm. of the ACM, vol. 45(2), 61–65, 2002.

[10] ISWC 2006. "OWL specification of the Semantic Web (SW) Topics Ontology". http://lsdis.cs.uga.edu/library/resources/ontologies/swtopics.owl

[11] R. Cyganiak, R. Delbru, H. Stenzhorn, G. Tummarello, S. Decker. "Semantic sitemaps: efficient and flexible access to datasets on the semantic web". *ESWC 2008* (Tenerife, Spain), LNCS 5021, 690–704.

[12] K. Alexander, R. Cyganiak, M. Hausenblas, J. Zhao. "Describing linked datasets". *LDOW 2009* (Madrid, Spain).

[13] F. Maali, J. Erikson, P. Archer. "DCAT catalog vocabulary". W3C Working Draft, 2012.

[14] J. Breuker, W. van de Velde. "CommonKADS library for expertise modeling: Reusable Problem Solving Components". IOS Press, 1994.

[15] G. Dromey. "Scaleable formalization of imperfect knowledge". AWCVS 2006 (Macao), 21–33.

[16] J.Z. Pan, L. Lancieri, D. Maynard, F. Gandon, R. Cuel, A. Leger. "Success stories and best practices". Deliverable D1.4.2v2 of KWEB (Knowledge Web), EU-IST-2004-507482.

[17] V. Presutti, A. Gangemi. "Content Ontology Design Patterns as practical building blocks for web ontologies". ER 2008 (Spain) http://ontologydesignpatterns.org.

[18] L. Dodds, I. Davis. "Linked Data patterns – a pattern catalogue for modelling, publishing, and consuming Linked Data". http://patterns.dataincubator.org/book/, 56 pages, 2011.

[19] B.K. Kahn, D.M. Strong, R.Y. Wang. "Information quality benchmarks: product and service performance". Communications of the ACM, vol. 45(4) 2002, 184–192.

[20] G. Mcdonald. "Quality indicators for Linked Data datasets". http://answers.semanticweb.com/questions/1072/quality-indicators-for-linked-data-datasets (2011).

[21] A. Flemming, O. Hartig. "Quality criteria for Linked Data sources". http://sourceforge.net/apps/mediawiki/trdf/index.php?title=Quality_ Criteria_for_Linked_Data_sources (2010).

[22] P.N. Mendes, C. Bizer, Y.H. Young, Z. Miklos, J.P. Calbimonte, A. Moraru. "Conceptual model and best practices for high-quality metadata". Deliverable 2.1 of PlanetData, FP7 project 257641 (2012) .

[23] C. Bizer. "Quality-driven information filtering in the context of web-based information systems". PhD dissertation (195 pages), Free University of Berlin, 2007.

[24] C. Mader. "Quality criteria for SKOS vocabularies". https://github.com/cmader/qSKOS/wiki/Quality-Criteria-for-SKOS-Vocabularies (2012).

[25] C. Fürber. "Data quality constraints library". http:// semwebquality.org/ documentation/primer/20101124/ (2010).

[26] A. Gómez-Pérez, F. Ciravegna. "SEALS EU infrastructures project – semantic tool benchmarking". http://www.seals-project .eu/ (2012).

[27] A. Hogan, A. Harth, A. Passant, S. Decker, A. Polleres. "Weaving the pedantic web". LDOW 2010 (Raleigh, USA).

[28] E. Pietriga, C. Bizer, D. Karger, R. Lee. "Fresnel: a browser-independent presentation vocabulary for RDF". ISWC 2006 (USA), LNCS 4273, 158–171.

[29] Ph. Martin. "Collaborative knowledge sharing and editing". IJCSIS vol. 6(1), 2011, 14–29. ISSN: 1646-3692.

Multimedia Content Identification Through Smart Meter Power Usage Profiles

Ulrich Greveler*, Peter Glösekötter‡, Benjamin Justus†, Dennis Loehr†
*Department of Communications & Environment, Rhein-Waal University of Applied Sciences, D-47475 Kamp-Linfort, Germany
ulrich.greveler@hochschule-rhein-waal.de
†Computer Security Lab, Münster University of Applied Sciences, D-48565 Steinfurt, Germany
{benjamin.justus, loehr}@fh-muenster.de
‡Department of Electrical and Computer Sciences, Münster University of Applied Sciences, D-48565 Steinfurt, Germany
peter.gloesekoetter@fh-muenster.de

***Abstract*—Advanced metering devices (smart meters) are being installed throughout electric networks in Germany (as well as in other parts of Europe and in the United States). Unfortunately, smart meters are able to become surveillance devices that monitor the behavior of the customers. This leads to unprecedented invasions of consumer privacy. The high-resolution energy consumption data which are transmitted to the utility company allow intrusive identification and monitoring of equipment within consumers' homes (e. g., TV set, refrigerator, toaster, and oven). Our research shows that the analysis of the household's electricity usage profile at a $0.5s^{-1}$ sample rate does reveal what channel the TV set in the household was displaying. It is also possible to identify (copyright-protected) audiovisual content in the power profile that is displayed on a CRT[1], a Plasma display TV or a LCD[2] television set with dynamic backlighting. Our test results indicate that a 5 minutes-chunk of consecutive viewing without major interference by other appliances is sufficient to identify the content.**

Our investigation also reveals that the data transmitted via the Internet by the smart meter are unsigned and unencrypted.

Our tests were performed on a sealed, operational smart meter used for electricity metering in a private home in North Rhine-Westphalia, Germany. Parameters for other television sets were obtained with an identical smart meter deployed in a university lab.

Keywords. Smart Meter, Data Privacy, Audiovisual Content, Smart Grid

I. Introduction

A smart meter is an electrical meter that records consumption of electrical energy at intervals and has the capabilities of communicating between a central server of its recorded information. The installation of smart meters at private homes is planned in Germany, as well as in EU in the near future. By 2020, the smart metering devices are supposed to replace 80% of the existing conventional meters. Smart metering is believed to be a crucial factor for the future availability of supply, energy efficiency and renewable energy[3]. From a consumer perspective, smart metering offers potential benefits such as: consumers by using a smart meter are able to view their detailed energy consumption data via a web-browser. The visualization of these data lets consumer to see into details how energy at home is used, therefore providing possibilities for devising energy saving strategies in view of their energy consumption habits. The energy company can also use the smart meter data for the purposes of infrastructure planning, network optimization and load balance checking. One also sees a trend towards IP-based communication as a common platform for smart meter applications[4].

Smart meter data contain consumer's personal information (see section IV). Also depending on the granularity of measurement and the resolution of data, we show in this paper that it is possible to deduce personal behavior of an individual in a private home. These behaviors include for example what TV channels, and which movies an individual has viewed in the course of a smart meter recording. In view of the concerns above, there are henceforth urgent calls for researchers to provide means of better protecting data transmitted by a smart meter.

II. Related Work

Even before the advent of smart meters, extensive researches have been done on techniques of non-intrusive load monitoring (NILM). Various NILM methods [12], [16] are introduced in order to glean into detailed energy consumption pattern in a household. Using these techniques, it turns out that a remarkable number of electric appliances in a private home can be identified by their load signatures with impressive accuracy. The same NILM techniques can be applied to analyze smart meter data in order to peek into household activities [17]. More recently, the authors of [6] claimed that they were able to discern video contents from electromagnetic interference (EMI) signatures produced by different TV sets.

There have been privacy concerns over the deployment and usage of smart meters [2], [13], [18] in U.S. and Europe, precisely because they can inadvertently leak detailed information

[1]Cathode Ray Tube

[2]Liquid Crystal Display

[3]*80% Smart Meter Adoption By 2020 Through EU Mandate*: Yahoo Finance Report from Sep 29, 2011 8:10 AM

[4]*Its Official: The Future of the Smart Grid Is IP*: By Katie Fehrenbacher, Sep. 7, 2010, 7:57 AM, on gigaom.com

about household activities.There are currently two approaches of implementing privacy preserving smart meter data analysis. The first approach relies on masking the meter readings. The actual meter reading is adjoined by a masking value, in such a way that an adversary can not recover individual readings. Yet, the sum of the masking values across meters sums to zero. This technique is introduced [10], [11] to compute metering aggregation over a network. And most recently following this line, [4] developed a scheme that ensures the property of differential privacy. The second approach relies on homomorphic encryption. The metering aggregation using this approach is discussed in [7], and the algorithm within also allows detection of leakage in electricity distribution. Furthermore, in [5], [3] are introduced protocols to privately derive and prove the correctness of bills. Recently, a billing protocol based on Pedersen commitments and a plug-in privacy components is introduced in [9]. Finally, in [15] are introduced solution of embedding Trusted Platform Module (TPM) in the smart Meter to obtain signed tariff data.

III. Experimental Results

Our investigation aims to answer the following questions: (1) What are the possible ways of obtaining and evaluating data coming from a calibrated smart meter? (2) What can be deduced from smart meter data regarding a person's TV watching habit in a private home? The experiments mentioned in this paper took place from August to November, 2011.

A. Hardware Background

The tested smart meter had been acquired from the company Discovergy GmbH (Heidelberg, Germany) after signing a private household contract. This calibrated smart meter is installed in a typical private house in the region North Rhine-Westphalia, Germany. After the installation, the new meter replaces the conventional meter which is manufactured by the German public utility company RWE AG.

The Discovergy product is based on the smart meter model manufactured by EasyMeter GmbH, Bielefeld (Electronic 3-phase meter Q3D-A1004 v3.03). The smart meter takes measurement at an interval of two seconds. All data are transmitted to the servers hosted by Discovergy. The customers are then able to access these data via a web-browser. Discovergy[5] claims in its contract complete data encryption for each smart meter equipped household.

IV. Data Transmission

The transmission of smart meter data to the Discovergy-Server is done through the TCP/IP protocol. The meter is directly connected with a LAN/DSL router and receives its dynamic IP address via the DHCP protocol. Contrary to the company claim, the smart meter data are not encrypted. The energy consumption data are saved in a textfile format, while being transferred to the central servers. Figure 1 shows a snapshot of a typical data transmission. The unencrypted data can be easily hacked out.

[5]www.discovergy.com

```
POST /api/w.html HTTP/1.1
Content-Type: application/x-www-form-urlencoded
Host:85.214.93.99
Content-Length:851

version=0.9&identity=[redacted]&msg=228601&values=[
{"meterdata":"00000285.9823514*kWh","tickdelta":"00000285.9822239*kWh","seconds":"399511319.61"},
{"meterdata":"00000285.9824793*kWh","tickdelta":"00000285.9823514*kWh","seconds":"399511321.61"},
{"meterdata":"00000285.9826075*kWh","tickdelta":"00000285.9824793*kWh","seconds":"399511323.61"},
{"meterdata":"00000285.9827358*kWh","tickdelta":"00000285.9826075*kWh","seconds":"399511325.62"},
{"meterdata":"00000285.9828636*kWh","tickdelta":"00000285.9827358*kWh","seconds":"399511327.62"},
{"meterdata":"00000285.9829915*kWh","tickdelta":"00000285.9828636*kWh","seconds":"399511329.62"},
{"meterdata":"00000285.9831196*kWh","tickdelta":"00000285.9829915*kWh","seconds":"399511331.62"},
{"meterdata":"00000285.9832476*kWh","tickdelta":"00000285.9831196*kWh","seconds":"399511333.62"}]
&now=399511335.65
```

Fig. 1. Captured communication between smart meter and server

In addition, none of the data are signed. The identity (highlighted in black in Figure 1) of any smart meter is immediately revealed when the data are being transmitted to the central servers and could be used by an attacker to send different power consumption data to the server.

A. Resolution of data presented to the customer

Discovergy offers a web browser based view on the power consumption profile. A java-script based application requests the data from the Discovergy server and offers the visualization of the profile[6]. A typical profile example can be seen in Figure 2.

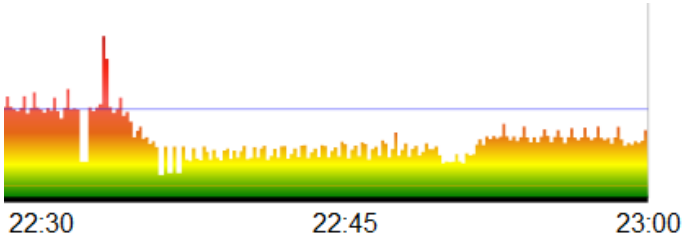


Fig. 2. Power profile visualized by Discovergy

An analysis of the script source code shows that the customer does not see the full resolution of the data (sampling rate at $0.5s^{-1}$). The data are consolidated by skipping absolute values (thus the arithmetic mean of several values is displayed). Moreover, a software bug regarding the time stamp parsing algorithm results wrong peaks and even negative peaks display. Such data would contradict the fact that the tested meter only submits monotonically increasing values (see captured communication in Figure 1). By correcting the software bug and downloading available meter data from the Discovergy server with a self-developed script we were able to visualize the complete data at various resolutions. Figure 3 depicts a small interval of data contained in Figure 2 (in the sub-timeframe 22.35h-22.50h).

B. Large Device Identification

We could verify the claims of other researchers [8], [13], [14]: electric appliances in a private home can be identified

[6]Note that this description reflects the state of the customer information portal *www.discovergy.com* during the test period (Aug. – Oct. 2011). It might be different at a later stage.

by their load signatures. In particular, we could identify the following household appliances: refrigerator, electric kettle, flow heater, light bulbs, energy-efficient lamps, bean-to-cup coffee machine, cooker hood, microwave oven, electric kitchen stove, washing machine, dishwasher, and the television set.

V. TV/Film Detection

A. Television Hardware

The first part of tests were performed on an home LCD television set in a household where the operational smart meter was installed. Liquid Crystal Display televisions use the display technology to produce colored images. Since the total amount of visible brightness of a picture is a combination of the backlighting and LCD shuttering, a technology dubbed *dynamic backlighting* is applied on modern LCD TVs to improve the contrast ratio[19]. While the shutters produce a contrast ratio of 1000:1, dynamic backlighting enhances this ratio up to 30000:1. The LCD TV power consumption is mainly influenced by the backlighting activities [1].

The experiment results presented in the following sections were obtained by using the household's Panasonic LCD television set[7]. Section V-H contains comparison results which use other TV models. The power consumption difference of a frozen white picture to a frozen black picture for this particular television was measured to be about 70 watts.

B. Power Consumption Prediction Function

The core of our content identification program is the power consumption prediction function. We explain below in details the construction of the function. The input of the function is the multimedia content, the output is power usage prediction as would displayed by a smart meter.

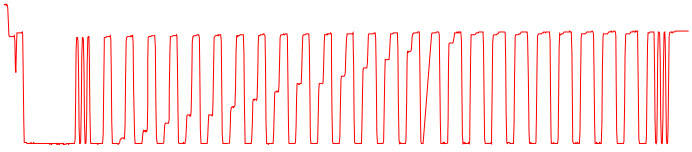


Fig. 3. Determination of b_{min}

The first step is to measure the power consumption for a series of pictures consisting of elementary shades. We use the additive RGB color notation with one byte (i. e. values 0–255) per red, green and blue portion. The sequence of pictures are then RGB 0-0-0, RGB 1-1-1, ..., RGB 255-255-255 that increase the brightness from black to white running over 254 shades of gray. Our observation shows that maximum power consumption is reached with rather dark pictures (e. g., RGB 32-32-32). But this also depends on the television user settings. For the rest of the paper, we denote this value by b_{min} which is the minimum brightness value that maximizes TV power consumption. A typical b_{min} value for the tested LCD TVs lies in the range $\{26, \ldots, 58\}$.

[7]Panasonic model number TX-L37S10E

$$n := 2 \text{ times (no. of frames per second)}$$

$$m_i := \begin{cases} 1 & \text{if } b_i > b_{min} \\ \frac{b_i}{b_{min}} & \text{otherwise} \end{cases} \qquad pp_k := \frac{1}{n} \sum_{i=nk}^{n(k+1)-1} m_i$$

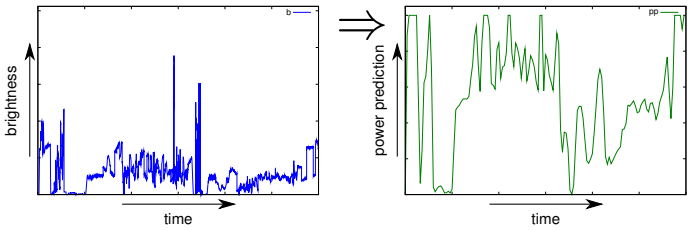


Fig. 4. Power prediction is computed on frame brightness values

Figure 3 shows one of the test runs we had performed in order to determine the value b_{min}. A sequence of pictures was shown: *black-white* (3 times) as a trailer to find the signal, then *black-(RGB-2-2-2)-white-black-(RGB-4-4-4)-white-black-(RGB-6-6-6)...* to see the increasing power consumption. One can then count the number of peaks until the gray picture (here: RGB-38-38-38) reaches maximum power consumption, i.e. becomes indistinguishable from white with regard to the power profile. At a later stage we did not need to run these tests anymore since we developed a script that performs content identification by automatic parameter detection.

The next step (shown in Figure 4) is to extract frames from the movie and determine the brightness of each frame. The mean value of the red, green and blue portion is calculated to be the frame brightness value b (or value b_i for a frame with index i). By assuming a linear function (suggested by the results of step one) we can then let the predicted power consumption m_i (for a frame with index i) to be at the TV set's maximum power consumption for all frames being brighter then (*RGB* b_{min}-b_{min}-b_{min}) and being equal to $(max - min)(b_{min} - b)$ for all frames with brightness $b < b_{min}$. To be more TV device independent we use a function with values from 0 (minimum power consumption) to 1 (maximum power consumption).

$$m_i := \begin{cases} 1 & \text{if } b_i > b_{min} \\ \frac{b_i}{b_{min}} & \text{otherwise} \end{cases}$$

As we obtained our experimental results with a smart meter operating on a two-seconds interval, we then calculate an average value of power consumption for a number of consecutive frames adding up to two seconds of a movie, e.g. 50 frames for a movie with a typical 25 frames per second (fps) rate.

$$pp_k := \frac{1}{n} \sum_{i=nk}^{n(k+1)-1} m_i$$

Our derived power prediction function does then give a predicted power consumption value after 2s ($k = 1$), 4s

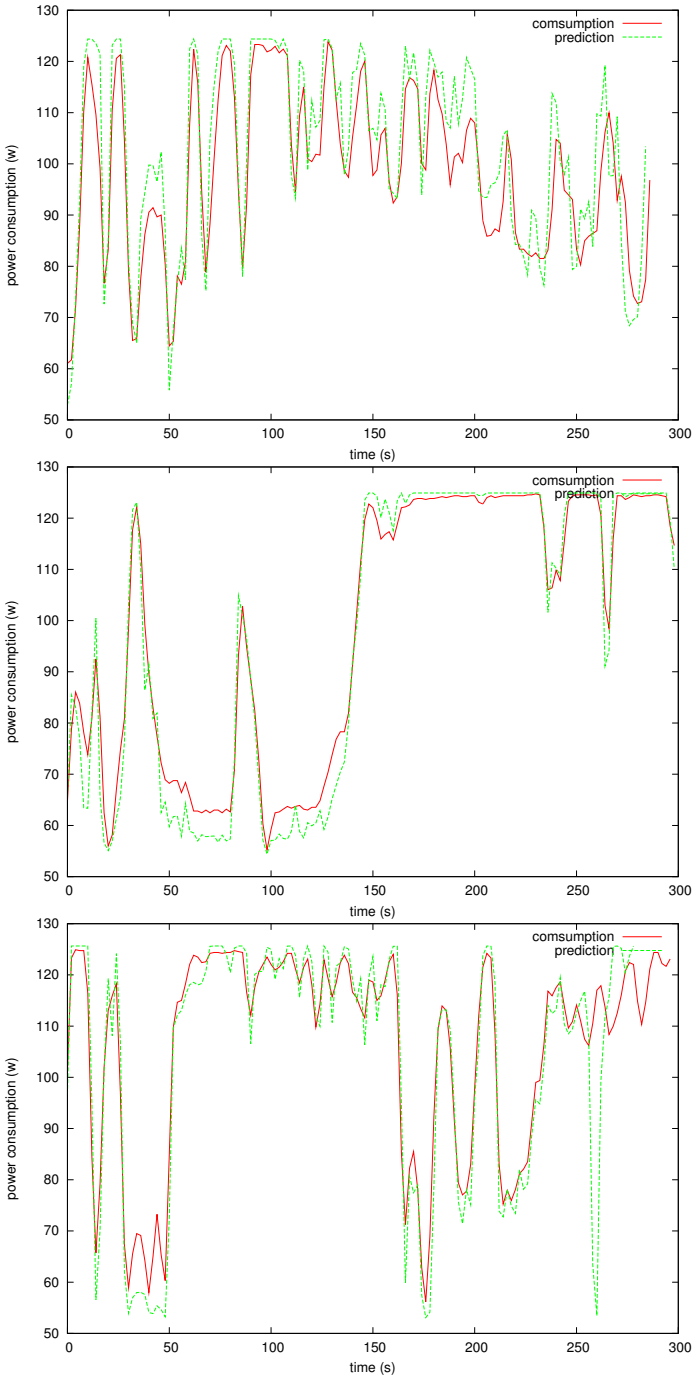


Fig. 5. power prediction vs. consumption: first 5 minutes of the movie Star Trek 11 (top), of episode 1, Star Trek TNG season 1, of the movie Body of Lies (bottom)

($k = 2$), 6s ($k = 3$), etc. This data can be correlated with any subsequent power profile data of the same length in order to search for the content.

C. Preliminary Analysis

To test our prediction function, we did a preliminary run on some films. We extracted first 5 minutes of each movie file, and then compared the actual power consumption against values produced by the prediction function. The movies we used for the test are:

1. Movie *Star Trek* (2009). Directed by J. J. Abrams. Release date: May 8, 2009.

2. Star Trek episode *Encounter at Farpoint* (1987). Directed by Corey Allen. Original air date: September 28, 1987.

3. Movie *Body of Lies* (2008). Directed by Ridley Scott. WarnerBros. Pictures. Release date: October 5, 2008

The actual power consumption was measured using a sealed operational smart meter while the films are playing on the household television set. No major appliances were operating during the measurements, only lights and stand-by consumption were active.

Figure 5 contains the experimental results. The green dotted curve is the prediction, and the actual power consumption data is plotted in red. We also calculated the Pearson product-moment correlation coefficients between the actual and predicted power consumption data. The correlation for the three movie events are 0.94, 0.98 and 0.93 respectively.

D. Corridor Algorithm

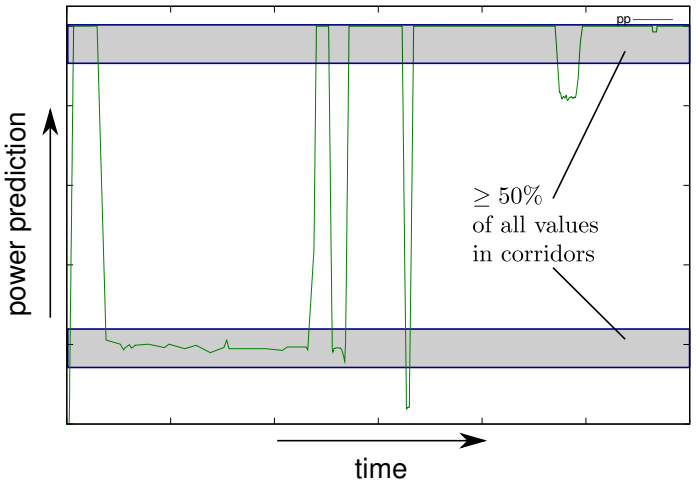


Fig. 6. Corridor algorithm discards chunk where more than half of values are found in distinct two corridors

During the experiments, we have noticed that the power consumption curve as observed by a smart meter oscillates in a normal household situation (without TV running) in a way that could lead to false positive identification of TV content. The reason for that is while searching for 5-minute chunks of movie files, if the chunk is for example showing a long dark scene, followed by a long bright scene (both scenes added exceed five minutes), it will correlate strongly with a power curve that reflects the switching-on of a simple electric appliance (e. g. a light bulb). So any curve jump phenomenon comes into the movie detection scenario (Picture 2 of Figure 5) showing a long bright scene could lead to false positive matching. To make movie load signature more distinguishable, it is desirable to eliminate possible false matches reflecting this effect during the analysis stage. For that purpose, we have developed a *Corridor Algorithm*. If too many values

of predicted or actual power consumption fall in one of two corridors, this movie-chunk will be discarded. Figure 6 shows a typical scenario, in which the green power curve is truncated within the corridors which are highlighted in gray. The parameters for the decision (threshold, corridor heights) are derived in section V-I.

E. Automatic Detection of b_{min}

In order to identify a broad range of video material, we have developed a script to detect the optimal b_{min} values for each video content played. For each possible b_{min} value ($= 0 \ldots 255$), the correlation between actual consumption and power prediction is calculated. Figure 7 contains a comparison between actual power consumption curve with predictions supported by various b_{min} values.

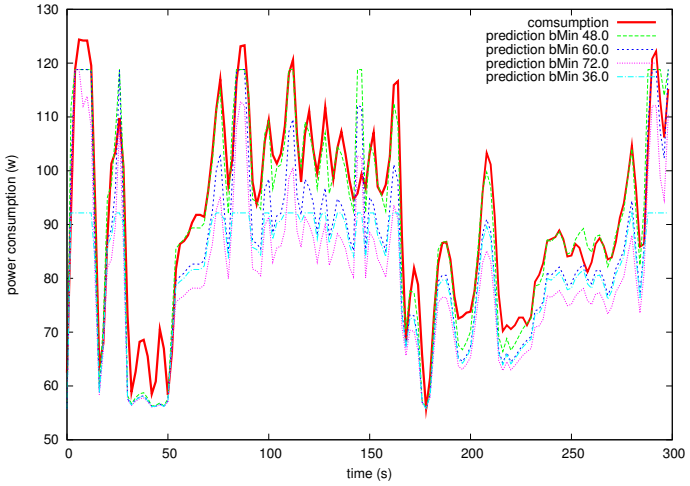


Fig. 7. Determination of b_{min}

F. Work-flow

This section describes the work-flow that are involved in the movie identification process. The Figure 8 illustrates all the steps involved. These steps are performed automatically by a software script we developed for the research being described in this paper and which could be regarded as a proof-of-concept for a forensic tool performing content identification on power consumption data.

The entire film is first divided into 5 minute chunks, and the brightness of each frame is calculated. The correlation value for the chunk is then computed using the predicted power values. The matches (correlation value is greater than 0.85 for a generic b_{min}-value) are further processed with a b_{min}-optimization algorithm and the corridor algorithm. The chunk is discarded if the threshold is reached. It should be noted that the identification process fails on some of the 5-minutes movie chunks due to either power disturbance or user interaction with the TV or the playing device. For a typical 90 minutes playback, we have $90/5 = 18$ blocks at disposal, so in a actual test there should be a good chance that at least two or three of these chunks *survive* other appliances' activities and can be found in the power curve matching.

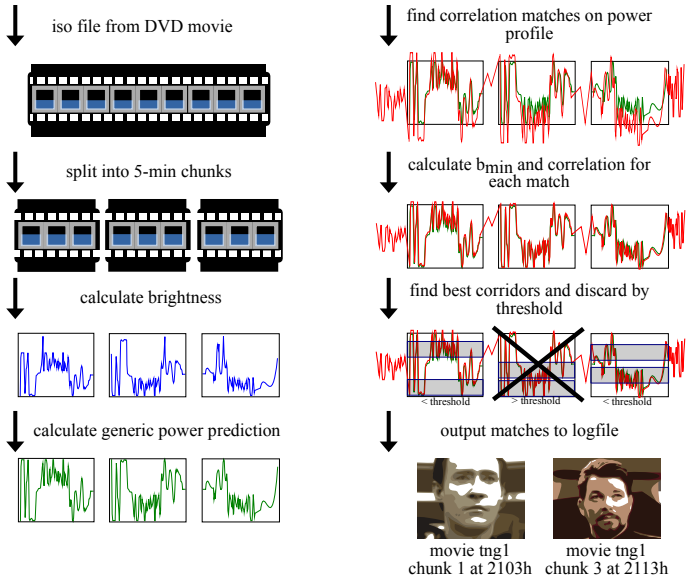


Fig. 8. Work-flow to detect chunks of a movie

G. Discoverable Video Material

During our experiments, some recorded television content such as German daily news[8] are not identifiable due to: lighting level of each content block consistently stays about the same level. This leads to an almost flat line in the power prediction curve for backlighted LCD TVs and the fluctuation of power consumption therefore can not be detected (detection is though still possible with CRT TVs: see Table I). Some other content such as the JAG[9] TV series has higher brightness level than typical b_{min} values: detection is harder when the shows is played on LCD TVs. Having gained some experiences with 653 content files and some days of recorded program broadcast, we could state that detection of movies produced for cinema projectors was almost always a feasible task while many TV studio productions (e. g. talk shows, news) are difficult or impossible to identify when played as recorded content. It is still rather easy to determine which TV broadcast consumers are watching since we only have to correlate the power curve to the content of a few dozen live stations and there is no need to search for a match along the time axis as the timing is provided by the program. The second or third rare dark scene (with brightness $b < b_{min}$) is then sufficient to identify the station on the power curve (idea: the first matching scene could be a coincidence: e. g. viewer interaction with TV teletext).

H. Other Television Models

Experiments described in the previous sections were performed on a home LCD TV[10] equipped with dynamic backlight enhancing technology. To support our claim that content

[8] *ARD Tagesschau*: daily German news broadcast at 8 p.m.
[9] Director: Donald P. Bellisario, air date: 1995 - 2005
[10] Panasonic model number TX-L37S10E

TABLE I
LIST OF TELEVISIONS

Manufactor	Model Nr. technology	$watt_{\min}$	$watt_{\text{diff}}$	b_{min}	correl. $M1$[12]	$M2$[13]	$M3$[14]	TV show[15]
Panasonic	TX-L37-S10E LCD	~ 45	~ 70.0	$26-58$	0.9599	0.9283	0.9487	< 0.85
LG	47LH4000 LCD	~ 65	~ 1.5	$25-84$	0.9458	< 0.85	< 0.85	< 0.85
Orion	TV32FX-100D LCD	~ 100	~ 3.0	$50-232$	0.8958	0.9402	0.9326	0.8989
Panasonic	TX-P50S-20E Plasma	~ 45	~ 160.0	$81-92$	0.8722	0.9510	0.8871	0.8933
Sony	KDL46EX-505 LCD	~ 170	–	–	< 0.85	< 0.85	< 0.85	< 0.85
Telefunken	Cinevision 32 CR tube	~ 60	~ 50.0	$58-153$	0.8833	0.9454	< 0.85	0.9283

identification is in general possible, we have performed experiments on other TV models as well.[11] We connected a out-of-box smart meter (meter Q3D from Easymeter GmbH) with specific TV sets.

For those tests not conducted with the operational smart meter, all data transmission took place directly between the smart meter and a connected notebook computer. We have played a total of 7 films and 2 TV shows. Table I contains the detailed test results relating to 3 movies and 1 TV show.

The successful test results affirm our belief that movie/TV content identification via fine-grained smart meter data is possible (with the exception of the Sony LCD test case). We also would like to point out that content identification using a cathode ray tube or a plasma display is also possible if very bright frames are taken into consideration. For some LCD TV sets not supporting dynamic backlighting (see sets *LG*, *Orion* in Table I) the power consumption difference between light and dark frames is rather small so content identification might become infeasible, especially if other appliances generate some noise to the smart meter data.

I. Determining the Optimal Values

To optimize the set of parameter values (minimum correlation that qualifies a matching content, corridor height, threshold for discarding a match), a series of test runs for varying parameters are performed (Table II). Starting from values suggested by the first experimental results, different values were subsequently tried. The number of false positive and correct identification (last two columns) are recorded. In conclusion, the combination $0.85, 50\%, 5\%$ (second last row) seems to produce the best identification rate. This means that the 5 minute chunks are compared to the power curve using a sliding window along the power axis and a preliminary match is declared when the correlation exceeds 0.85. The match is discarded if more than 50% (threshold) of the values fall in two corridors each having a height of 5% of the whole interval. Two optimal corridors maximizing the corridor coverage are to be identified for each match. The discarding is done twice: On the predicted power consumption values and on the measured power consumption values.

The goal of the elimination process is to prevent false positive matches but it also leads to discarding of about half of the correct hits (Table II shows the results for 12 movie chunks not being discarded by predicted power values). Since a movie consists of 18 or more 5 minute chunks, this discarding procedure is applicable in a real-life scenario of content identification. The parameter combination $0.85, 50\%, 5\%$ provides the identification of 11 out of the 12 chunks while one false positive match was logged. We used a collection of 653 content files to search for content matches.

[11]The authors wish to thank graduate student Stephan Brinkhaus BSc. who conducted various tests with the smart meter on several TV sets and other appliances.

TABLE II
CONTENT IDENTIFICATION OF 12 MOVIE CHUNKS WITH DIFFERENT PARAMETERS

Correlation	Thresh.	Height	PC Codr	PC Height	False Positive	Identification
0.9	0.8	0.10	0.8	0.10	5	9
0.85	0.8	0.10	0.8	0.10	12	11
0.8	0.8	0.10	0.8	0.10	69	12
0.9	0.7	0.10	0.7	0.10	2	9
0.85	0.7	0.10	0.7	0.10	4	11
0.8	0.7	0.10	0.7	0.10	34	12
0.9	0.6	0.10	0.6	0.10	0	7
0.85	0.6	0.10	0.6	0.10	0	8
0.8	0.6	0.10	0.6	0.10	1	8
0.9	0.6	0.05	0.6	0.10	0	8
0.85	0.6	0.05	0.6	0.10	1	9
0.8	0.6	0.05	0.6	0.10	5	9
0.9	0.6	0.05	0.6	0.05	2	9
0.85	0.6	0.05	0.6	0.05	6	11
0.8	0.6	0.05	0.6	0.05	39	11
0.9	0.5	0.05	0.5	0.05	0	7
0.85	0.5	0.05	0.5	0.05	0	8
0.8	0.5	0.05	0.5	0.05	3	8

J. False Positives with Other Appliances

In order to get some consolidated findings regarding false content identification, we used our scripted content identification work-flow (depicted in Figure. 8) to search for content in several 24h-periods, in which power metering data are concurrently generated by different household appliances. Four persons were living in the household and using the appliances. We used our available set of 653 content files – split into 5-minute chunks – to search for film material. We count a (false positive) hit for every match having a correlation of at least 0.85, and there are 35.5 hits per 24 hours (see Figure 9 for example hits' log entries).

```
INFO  First correlation discard threshold: 0.0
INFO  Second correlation discard threshold: 0.85
INFO  Corridor discard threshold: 0.5
INFO  Corridor height: 5
INFO  Power consumption corridor discard threshold: 0.5
INFO  Power consumption corridor height: 5
INFO  Analyze log file "discovergy-Raw-2011.11.19_0100-2011.11.19_2359.csv"
INFO  *(_csv_5Min.Saw.6.1080p.mkv.csv) at 07:14:50
        cor =          0.8582734470331655      bMin =          40.0
        delta =        2.160000520199958       distance =      0.40085441550795814
        corridor =     0.4533333333333333      pcCorridor =    0.31    [...]
INFO  *(_csv_5Min.Spaceballs.mkv.csv) at 11:06:43
        cor =          0.8554506191367586      bMin =          28.0
        delta =        316.08000034434         distance =      57.05935049506766
        corridor =     0.42                    pcCorridor =    0.3333333333333333
INFO  *(_csv_5Min-Jackie.Chan.-.Action.Hunter.avi.csv) at 12:54:07
        cor =          0.8794719071839578      bMin =          48.0
        delta =        4239.1799992422         distance =      466.5877682097706
        corridor =     0.25333333333333335     pcCorridor =    0.36    [...]
```

Fig. 9. Log file clipping showing false positive matches on other appliances

Analyzing the hits shows that these can easily be identified as false positive matches because the power curve does obviously not reflect television operation. See Figure 10 as an example showing such a false positive match: the power consumption difference of more than 4000 watts is too high

for being generated by a TV set and the curve shape does not obey the shape of prediction.

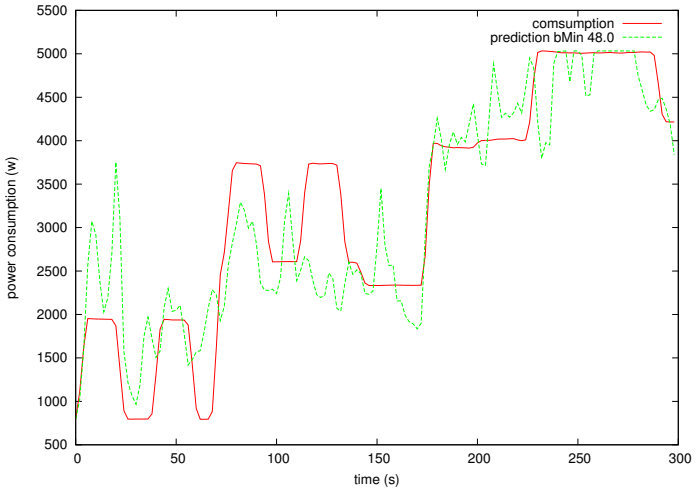


Fig. 10. Example of a false positive match

To avoid time-consuming manual discarding of the hits, a straightforward strategy to identify content would be to count only findings consisting of more than one match of a 5-minute chunk. A log entry showing two corresponding chunks of the same content (like the example of Figure 8: chunk 1 at 21:03h and chunk 3 at 21:13h) ruled out false positives on other appliances during the experiments. Note that we only used 653 content files for our experiments; a forensic investigator who is searching for copyright-protected content of all movies and TV productions ever been produced might have to solve a more challenging false-positive-problem. We did not have sufficient content files to reach a proper assessment on the feasibility of a forensic software.

VI. Conclusion

Smart meters are able to become devices that monitor the behavior of their customers. The personal privacy invasion is obvious if the smart meter data are available to malicious parties or being used by members of the same household to spy on each other.

A new generation of smart meters generating high-resolution energy consumption data could henceforth cause new potentials concerns regarding consumers' privacy sphere. We have demonstrated that particular information available on appliances in the household via its detailed power profile allow a fine-grained analysis of the appliance's behavior. Taking measurements at an interval of two seconds is sufficient to enable the identification of a television program or audio-visual content if favorable conditions are in place (e.g., no major interference of other appliances for minutes long). Our research has shown that the electricity usage profile with a $0.5s^{-1}$ sample rate leads to a *invasion* into a person's private sphere regarding his TV watching habits. Five minutes of consecutive playing of a movie is in many cases sufficient to identify the viewed content by analyzing the smart meter power consumption data. While we did not have sufficient content files to generate affirmative statistical evidences that can lead to a forensic software that will police every copyright infringement material. Our paper shows that there is at least a major privacy issue regarding content identification via a smart meter.

Our investigation also reveals that a smart meter transmits data via the Internet unsigned and unencrypted. This is a major throwback in data integrity and consumer privacy. This technical flaw could be attributed to the startup nature of the installer company who is in a phase of service development and product quality definition. It nevertheless proves that a minimum regulatory requirement regarding smart meter data protection standards need to be defined and fulfilled, before a meter becomes fully operational and capable of preserving a user privacy.

References

[1] The basics of tv power. http://reviews.cnet.com/green-tech/tv-power-efficiency/, April 2010.

[2] Researchers analyze smart meter data. http://www.spiegel.de/netzwelt/netzpolitik/0,1518,787629,00.html, September 2011.

[3] A. Rial and G. Danezis and M. Kohlweiss. Differential private billing with rebates. Technical Report MSR-TR-2011-10, Microsoft Research, Feburary 2011.

[4] Gergely Ács and Claude Castelluccia. Dream: Differentially private smart metering. *CoRR*, abs/1201.2531, 2012.

[5] G.Danezis A.Rial. *Privacy-Preserving Smart Metering*, MSR-TR-2010-150.

[6] Miro Enev, Sidhant Gupta, Tadayoshi Kohno, and Shwetak N. Patel. Televisions, video privacy, and powerline electromagnetic interference. In *ACM Conference on Computer and Communications Security*, pages 537–550, 2011.

[7] Flavio D. Garcia and Bart Jacobs. Privacy-friendly energy-metering via homomorphic encryption. In J. Cuellar et al., editor, *6th Workshop on Security and Trust Management (STM 2010)*, volume 6710 of *Lecture Notes in Computer Science*, pages 226–238. Springer Verlag, 2010.

[8] G.W. Hart. Nonintrusive appliance load monitoring. *Proceedings of the IEEE*, 80(12):1870–1891, 1992.

[9] Marek Jawurek, Martin Johns, and Florian Kerschbaum. Plug-in privacy for smart metering billing. In *PETS*, pages 192–210, 2011.

[10] K. Kursawe. Some Ideas on Privacy Preserving Meter Aggregation. Technical Report ICIS–R11002, Radboud University Nijmegen, January 2011.

[11] Klaus Kursawe, George Danezis, and Markulf Kohlweiss. Privacy-friendly aggregation for the smart-grid. In *PETS*, pages 175–191, 2011.

[12] H. Lam, G. Fung, and W. Lee. A novel method to construct a taxonomy of electrical appliances based on load signatures. In *IEEE Transactions on Consumer Electronics*, 2007.

[13] Andres Molina-Markham, Prashant Shenoy, Kevin Fu, Emmanuel Cecchet, and David Irwin. Private memoirs of a smart meter. In *2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings (BuildSys 2010)*, Zurich, Switzerland, November 2010.

[14] Klaus Mueller. Gewinnung von Verhaltensprofilen am intelligenten Stromzähler. *Datenschutz und Datensicherheit - DuD*, 34:359–364, 2010. 10.1007/s11623-010-0107-2.

[15] Ronald Petrlic. A privacy-preserving concept for smart grids. In *Sicherheit in vernetzten Systemen: 18. DFN Workshop*, pages B1–B14. Books on Demand GmbH, 2010.

[16] A. Prudenzi. A neuron nets based procedure for identifying domestic appliances pattern-of-use from energy recording at meter panel. In *IEEE Power Engineering Society Winter Meeting*, 2002.

[17] E.L. Quinn. *Privacy and New Energy Infrastructure*. Available: http://papers.ssrn.com/sol3/papers.cfm?abstract id=1370731, 2009.

[18] Stephan Renner. *Smart Metering und Datenschutz in Oesterreich*, DuD 2011.

[19] Robert Scott West, Huub Konijn, Simon Kuppens, Nicola Pfeffer, Quint van Voorst Vader, Yourii Martynov, Tewe Heemstra, and Jan Sanders. Led backlight for large area lcd tvs. In *10th International Display Workshops (IDW)*, 2003.

Sparse Matrix Approximation Technology for Concept Decomposition on Large Datasets

Wen Li and Chi Shen
Division of Computer Science, Kentucky State University,
Frankfort, KY 40601, USA

Abstract

In information retrieval technology, text documents are modeled as term-document matrices in which terms are considered as rows and documents are considered as columns. Those matrices are usually highly dimensional and sparse. Take LISA as an example, it has more than 6000 dimensions and about 0.21% nonzero entries. To reduce the high dimensions, one of the various dimensionality reduction methods, concept decompositions to approximate the matrix of document vectors has been developed in [5]. However the numerical computation is expensive, as an inverse of a dense matrix formed by the concept vector matrix is required. A class of multistep sparse matrix strategies for approximating concept vector matrix in [3] has shown the efficiency in terms of computational and memory costs for small datasets. This motivates us to test the robustness of this algorithm and apply it on two large datasets LISA and NPL. In our numerical experiments, we continue to see the advantage of this approach in terms of computational costs and storage costs while maintaining the query result at an acceptable accuracy level.

Keywords: term-document matrix, concept decomposition matrix, multistep sparsity pattern, sparse matrix approximation, least-squares

1 Introduction

Nowadays people are overwhelmed by information from various sources. With large sets of documents increasing rapidly, being able to efficiently utilize this vast volume of new information and service resource presents challenges to computational scientists. Most knowledge is recorded in text documents and text documents are usually modeled as a term-document matrix which has high dimensional and sparse vectors, which can be thousands of dimensions and 95% to 99% of sparsity [5]. Various methods have been developed in order to extract high-quality data from the text documents. Among them, Latent Semantic Indexing (LSI) with the low-rank Singular Value Decomposition (SVD) of the term-document matrix have attracted a lot of researchers and been studied for decades. However, the practical success does not cover all the disadvantages of this method: it cannot fully interpret the low-rank approximation [1]and not useful for large text documents whose contents are not homogeneous [2, 10]. Concept decomposition for text data has been first proposed by [5]. And concept decompositions are introduced to approximate the sparse text documents and it is claimed that the errors produced by the approximation method can be compared to best results of text matrix truncated by singular value decomposition [5]. And also, the method showed that the concept vectors are sparse and close to be orthogonality. The approach improves the interpretation of the low-rank approximation.

However, when carrying out the concept matrix decomposition directly, an inverse matrix formed by the concept vector matrix, which is a dense matrix, consumes large CPU time and computational power at query stage [3]. The traditional straight forward computation method of the inverse concept vector matrix cannot fulfill the requirement of efficiency.

Multistep Sparse Matrix Strategies have been proposed in [3]. In this approach, a sparse matrix is computed to approximate the projection matrix in a straightforward way on the basis of the ideas from [4] and [2].

Shen et al made use of sparse inverse technologies in the preconditioning field for concept matrix decomposition approximation to originate non-square sparse matrices. In [3], the approach of a class of multistep sparse matrix strategies has been tested on the small text datasets such as CISI, CRAN and MED and the result demonstrated that its performance in accuracy retrieval can be better than the methods introduced in [3] while using less CPU time and memory space in the decomposition matrix computation [3].

	doc_1	doc_2	...	doc_n
Aaar	0	1	...	0
Aba	1	0	...	1
Aci	0	0	...	1
Aeg	0	1	...	1
		...		

Figure 1: Term-document matrix

In order to better understand the behavior and efficiency of the multi-step sparse matrix strategies, in this paper, MSSP algorithm has been tested to two famous large datasets, LISA and NPL. The test results from our approach are comparatively more advantageous than the concept projection method in terms of memory cost. The MSSP can achieve 96.3% of the precision achieved by CPM by using only 49.8% storage memory of it.

The paper is organized as follows: Section 2 describes some basic concepts underlying this research; The Multistep Sparse Strategy Process (MSSP) are discussed in Section 3; The information of datasets LISA and NPL as well as the way of preprocessing the datasets used in this research are given in Section 4; Numerical experiments are presented in Section 5; Section 6 is our conclusion and the ideas for future work.

2 Concept Decomposition and Approximate Least Square Based Retrieve Procedure

2.1 Document Clustering and Concept Decomposition

In information retrieval, text documents are modeled as a term-document matrix whose rows are the terms and columns are document vectors. Fig. 2.1 provides a glimpse of the term-document matrix.

The words .Aaar., .Aba. and so on are the terms used in the documents doc_1, doc_2, and so on. If a term has been used in one document, the number "1" will be marked in the corresponding box. The number "0" will be marked if the term is not in the document.

Suppose we are given the set of document vectors

$A_{m\times n} = [a_1, a_2, \ldots, a_n]$, where $a_j, (j = 1, 2, \ldots, n)$ is the jth document vector of m dimension in the collection. Usually, the term-document matrices are high dimensional as each data set may contain thousands of words and are very sparse, with 1%-5% or less terms in each document.

We first partition the documents into k disjoint clusters $\{\pi_j\}_{j=1}^k$ by using k-means or other clustering algorithms such that

$$\bigcup_{j=1}^{k} \pi_j = \{a_1, a_2, \ldots, a_n\} \text{ and } \pi_j \cap \pi_i = \phi \text{ if } j \neq i.$$

For each fixed j, $1 \leq j \leq k$, the centroid vector of each cluster is defined as

$$\widetilde{c_j} = \frac{1}{n_j} \sum_{a_i \in \pi_j} a_i,$$

where $n_j = |\pi_j|$ is the number of documents in cluster π_j. The centroid vectors are normalized in following

$$c_j = \frac{\widetilde{c_j}}{\| \widetilde{c_j} \|}, \ \ j = 1, 2, \ldots, k.$$

An intuitive definition of the clusters is that, if $a_i \in \pi_j$, then

$$a_i^T c_j > a_i^T c_l \quad \text{for} \quad l = 1, 2, \ldots, k, \quad l \neq j,$$

i.e., documents in π_j are closer to its centroid than to the other centroids. The centroid vectors are also called *concept vectors* [5]. The *concept matrix* can be defined as an $m \times k$ matrix such that, $C = [c_1, c_2, \ldots, c_k]$. The concept matrix C that has rank $m \times k$ is still a sparse matrix. For any partitioning of the document vectors, we can define the corresponding concept decomposition $\tilde{A}$ of the term-document matrix A as the least squares approximation of A onto the column space of the concept matrix C. $\tilde{A} = C\tilde{M}$, where $\tilde{M}$ is a $k \times n$ matrix that is to be determined by solving the following least squares problem

$$\tilde{M} = \arg\min_M \|A - CM\|_F^2. \tag{1}$$

It is well-known that problem (1) has a closed-form solution, i.e.,

$$\tilde{M} = (C^T C)^{-1} C^T A. \tag{2}$$

2.2 Retrieval Procedure and Approximate Least Squares Based Strategies

As indicated in [3], the following equation is used to compute the retrieval of the concept projection matrix (CPM) $\tilde{A}$ for a query vector q.

$$r^T = q^T \tilde{A} = q^T C (C^T C)^{-1} C^T A, \tag{3}$$

where r^T is the ranking vector. The equation above is not a good way of computing the retrieval result since it is required to compute the resulted matrix from $(C^T C)^{-1}$ and it is likely to be a dense matrix.

To make the retrieval procedure in Eq. (3) more efficiently, a sparse matrix approximation technique based on the static and dynamic sparsity pattern strategies has been studied in [3]. In this method, a sparse matrix M is computed to solve the least squares problem (1) approximately.

As indicated in [3], compared with the direct computation of CPM, the approaches of finding out M use less than 5% memory storage and obtain competitive retrieval accuracy.

3 Multistep Sparse Matrix Strategies

3.1 Computational Ideas

As it is demonstrated in [3], a sparse matrix M with certain constraints is constructed by solving the following equation:

$$f(M) = \min_{M \in \mathcal{G}} \|A - CM\|_F^2, \tag{4}$$

$\mathcal{G}$ is a set of sparsity pattern of M, and the minimization problem of the equation (4) can be reduced into the smaller problems which can be represented by the following equation:

$$\|A - CM\|_F^2 = \sum_{j=1}^{n} \|(A - CM)e_j\|_2^2 = \sum_{j=1}^{n} \|a_j - Cm_j\|_2^2, \tag{5}$$

where a_j and m_j are the jth column of the matrices A and M, respectively. (e_j is the jth unit vector.) The minimization problem represented in equation (5) is the same as solving the functions as follows:

$$\|Cm_j - a_j\|_2, \qquad j = 1, 2, \ldots, n, \tag{6}$$

with certain restrictions placed on the sparsity pattern of m_j. In other words, each column of M can be computed independently. This certainly opens the possibility for parallel implementation.

There are a variety of methods available to solve the small least squares problem (6). One of methods is to approximately decompose matrix M, which minimizes $\|CM - A\|_F$ for the given sparsity pattern.

3.2 Multistep Sparse Pattern Strategies

Since the matrix A is not a square matrix, it is not easy to compute the M in the preconditioning field. The dimensions of matrices C and M are not the same. In another word, C is an $m \times k$ matrix while M is a $k \times m$ matrix. The multistep sparse strategies developed in [6] cannot be applied to our case directly. We approach the problem for non-square matrix minimization in another way.

Suppose a simple and cheap sparsity pattern is chosen for M_1. M_1 is computed approximately for the equation $CM = A$ by using our modified sparse approximate inverse techniques developed in [3]. If CM_1 is not close enough to A, another approximate sparse matrix M_2 can be obtained from the equation $CM = A - CM_1$ such that $CM_2 \approx A - CM_1$. We sum the two matrices M_1 and M_2, such that $C(M_1 + M_2) \approx A$. Generally, $C(M_1 + M_2)$ should be closer to A than CM_1 does. This procedure can be repeated for a few times to obtain a sequence of sparse matrices $M_1, M_2, ..., M_l$ such that $C(M_1 + M_2 + ... + M_l) \approx A$.

In general, the higher accuracy the matrix M is, the more dense it is. Hence the computational cost also may increase. In our approach, since we sum the matrices at each step, it shouldn't have many new fill-ins.

The most important thing is to define the sparsity pattern for M_1. Note that the concept matrix C describes the relationship between the term vectors and the concept vectors. If a term is related to a concept vector, this relationship may be maintained in the approximate decomposition matrix in some sense. The sparsity pattern strategy is described below.

From the equation $CM_1 = A$, we know that $c_i m_j^{(1)} = a_{ij}$, where c_i is the ith row of C, $m_j^{(1)}$ is the jth column of M_1. The sparsity pattern of $m_j^{(1)}$ is given in this way: If a_{ij} is the largest entry in the jth column of A, the sparsity pattern of $m_j^{(1)}$ is the same as that of c_i. Here we use small matrices to illustrate our ideas. Suppose the three matrices are $C_{4\times3}$, $M_{1_{3\times5}}$ and $A_{4\times5}$. The pattern of $CM_1 = A$ is depicted by the Eq (7).

$$\begin{pmatrix} x & 0 & x \\ 0 & x & 0 \\ x & x & 0 \\ 0 & x & 0 \end{pmatrix}_{4\times3} \begin{pmatrix} - & - & - & - & - \\ - & - & - & - & - \\ - & - & - & - & - \end{pmatrix}_{3\times5} = \begin{pmatrix} 0 & x & 0 & x & 0 \\ 0 & 0 & 0 & x & x \\ x & x & 0 & 0 & 0 \\ x & 0 & x & 0 & 0 \end{pmatrix}_{4\times5} \tag{7}$$

Here, "x" denotes nonzero entry, "-" denotes undefined pattern. We determine the sparsity pattern of M_1 column by column. First, find the largest entry in each

column of A, suppose they are a_{31}, a_{12}, a_{43}, a_{14}, and a_{25} in Eq (7). Then the sparsity pattern of $m_1^{(1)}$, is the same as that of c_3 and the sparsity pattern of $m_2^{(1)}$ is the same as that of c_1, $m_3^{(1)}$ has same sparsity pattern of c_4 and $m_4^{(1)}$ is the same as that of c_1. Finally we have the sparsity pattern of M_1 like this: $\begin{pmatrix} x & x & 0 & x & 0 \\ x & 0 & x & 0 & x \\ 0 & x & 0 & x & 0 \end{pmatrix}$.

For the sparsity pattern of matrix M_2 at the second step, we consider the equation $CM_2 = A - CM_1$. We first sort non-zero entries in each column of $A - CM_1$. Then we go through all the entries from the largest to the smallest. In order to have different sparsity patterns from that of M_1, if the position of this largest entry has been used in matrix M_1, this entry will continue to be used in the sparsity pattern in M_2, but it is not considered as a new fill-in. We then check the second largest entry in the given column of $A - CM_1$. If the second largest entry hasn't been used in M_1, it will be selected as one new fill-in and used to determine the pattern of M_2 in the given column. A parameter ne is used to control the number of new fill-ins for each column of M_i at the ith step. So in general the matrix M_i is denser than M_{i-1}. We had conducted some numerical tests for this strategy. With a few new fill-ins for the matrix M_2, the 2-norm residual of $\|A - C(M_1 + M_2)\|_2$ reduces much more than that of $\|A - C(M_1 + M_2)\|_2$ with no more new fill-ins in M_2. That means the denser matrix $M_1 + M_2$ may hold more information. This multistep procedure may be repeated for a few times as needed.

The single-step approach is identical to the approach of SSP (Static Sparse Pattern approach) developed in [3].

4 Datasets

We apply this strategy to the two large datasets LISA and NPL.

The LISA collection is a test collection based on the 1982 Library and Information Science Abstracts database. The information provided by LISA only includes the title and abstract fields of 6004 documents. In the collection, there are 35 natural language queries with relevance judgment obtained from the students in 1983. The whole LISA dataset can be downloaded from the website:
http://ir.dcs.gla.ac.uk/resources/test_collections/lisa/.
LISA is a large dataset with a feature: the number of the terms is more than the number of documents, which is denoted as $m > n$.

NPL, also known as VASWANI, is another famous large dataset for text mining research test. Just like LISA, there are queries and relevance documents in the dataset package. The dataset can be downloaded from the website: http://ir.dcs.gla.ac.uk/resources/test_collection. Compared with other datasets like CRAN, CISI, and LISA, the terms of NPL are less than the number of documents, which is $m < n$. In our case, by manually comparing the terms' list provided in the dataset package named term-vocabulary, many noisy words have been moved into the English stoplist created for preprocessing NPL and finally 11380 terms are kept for the research here. The terms' numbers of NPL in [2, 7, 8] are much less than the number in our research. But since our research target is to test our algorithm on large size of dataset which is expected to contain thousands of documents and terms and the dataset of NPL we created still has the feature that is $m < n$, so we decided to use the current NPL in our research.

The tool we used in our research is a free software named as Text-to-Matrix Generator (TMG) [9]. We use this tool to convert the text files into numerical matrices for further processing. The code of this software is written in MATLAB programming language. So when running the program, MATLAB platform is required.

In this work, the SMART's "English stoplist" downloaded from the website: http://web.eecs.utk.edu/research/lsi/ has been used for removing the stop words since it is widely used by many researchers and it contains more words than the stoplist provided with TMG (there are 571 in the SMART and 439 in the TMG). For LISA dataset, the "English stoplist" has been directly applied to the dataset without any change, while for NPL the stoplist has been combined with some other manually removed words to form a special stoplist just for NPL.

The information about the databases are given in Table 1.

Table 1: The information of datasets

Dataset	Matrix size(terms×docs)	No. of queries
LISA	17482×6004	35
NPL	11380×11429	93

5 Numerical Experiments

In this section, we conducted some experiments on LISA and NPL to compare the performance of our proposed algorithms MSSP and CPM method that is based on the

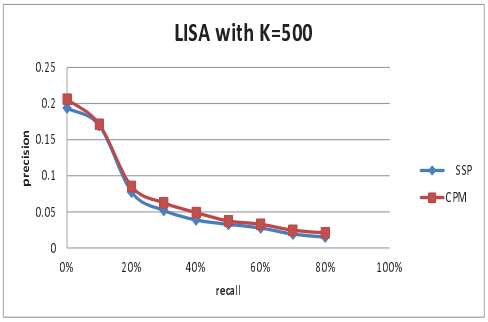


Figure 2: Precision-recall for LISA with $k = 500$ and single step

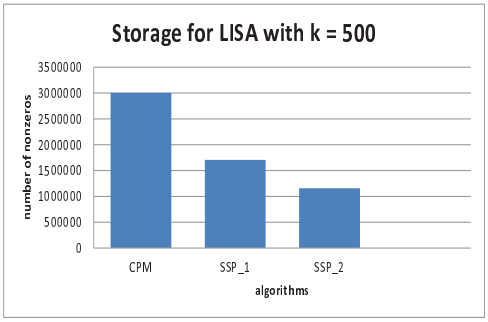


Figure 3: Storage costs for LISA with $k = 500$

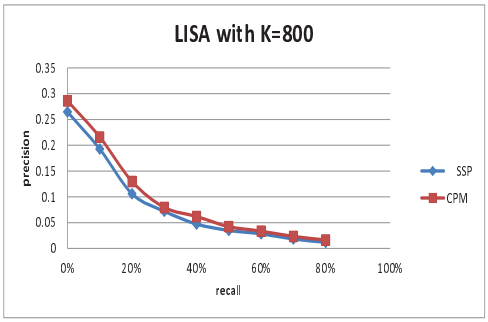


Figure 4: Precision-recall for LISA with $k = 800$ and single step

straightforward implementation by solving Eq.(2).

The metrics used to evaluate the validation of the algorithm is precision-recall curves. Precision is the proportion between the number of retrieved relevant documents and the total number of retrieved documents while recall is the proportion between the number of retrieved relevant documents and the total number of relevant documents. The precision of the queries of each dataset is represented at various recall values from $10\%, 20\%, \ldots, 90\%$.

In the experiments, the number of steps adopted in the algorithm can influence the sparsity of the matrix M. Normally, the more steps are used, the denser the matrices would be. However, denser approximate matrix does not mean higher precision result. Besides, the threshold upon which the number of entries can be decided is also important to the construction of M. Take the NPL dataset 500 clusters as an example. When using 2 steps in the algorithm, if the threshold is equal with 0.003 the number of the nonzero entries of M is 1372746 while the threshold is 0.005 the number is 1285269. The denser the matrix is, the more information can be hold by the matrix and it takes larger portion of CPU resources to compute the matrix. However, the precision is not depending on the number of nonzero entries in the M: the denser matrices may have poorer precisions.

In all the following tests, "SSP" denotes single step strategy; when use 2 steps, we label it as "MSSP_2"; when 3 steps are adopted, we label it as "MSSP_3".

We first test the precision-recall curves and storage costs for the LISA databased with various number of clusters, $k = 500$ and 800.

In Fig. 2, the LISA dataset is clustered into 500. We have been trying for different thresh values (0.3 and 0.003) to see if denser M results better performance. Their storage costs is presented in Fig. 3. They have similar performance shown in Fig. 2. From above results, we observe that compared with the single-step strategy, the multistep strategy can produce higher precision result. This is because the matrix constructed by MSSP may hold more information than the one constructed by SSP. However, as we have mentioned, denser approximate matrix may not always lead to better performance. Besides, the thresh value is very essential during the construction of the M, so we use different threshold values to adjust the information held in M.

We use large number of cluster, $k = 800$, to evaluate its performance. precision results are very promising for both SSP and MSSP methods. We present MSSP result in Fig. 5 and SSP is slightly worse than MSSP. The storage costs are given in Fig. **??**. We see that by using 49.8% resources required by the CPM, the MSSP can achieve nearly 96% of the precision achieved by CPM.

The tests on LISA dataset with different k have shown the advantages of the multistep algorithms over CPM method.

We then take tests on NPL dataset and we have mixed results. The two figures, Fig. 6 and Fig. 7 are the precision-recall curves of NPL when it has been grouped

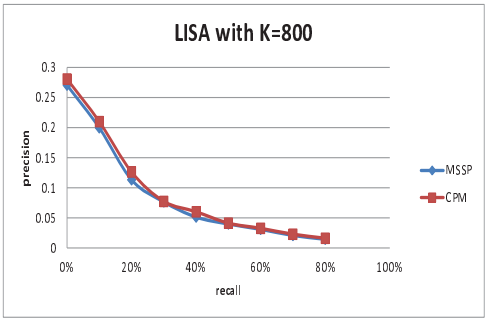


Figure 5: Precision-recall for LISA with $k = 800$ and 2 steps

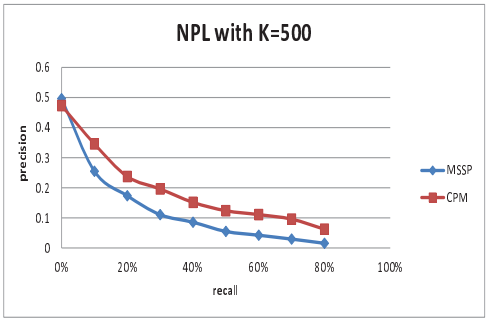


Figure 7: Precision-recall for NPL with $k = 500$ and 3 steps

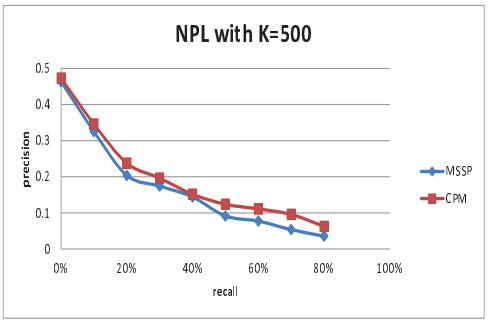


Figure 6: Precision-recall for NPL with $k = 500$ and 2 steps

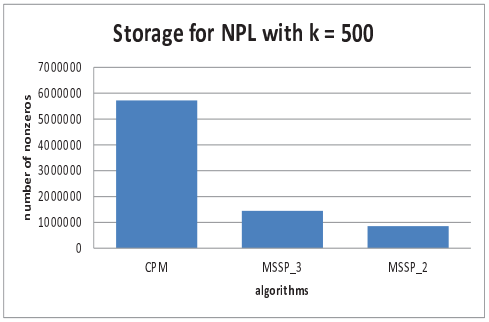


Figure 8: Storage costs for NPL with $k = 500$

into 500 clusters. From the curves, we can observe the results for MSSP in Fig. 7 is much better that in Fig. 6, although the nonzero entries in Fig. 6 are more than they are in Fig. 7. See the storage costs in Fig. 8. So, from this example, we can further conclude that in MSSP the denser approximate matrix M may not achieve better performance due to potential noisy data that are captured in the denser matrix.

6 Conclusion

The result of this study further proves the multistep sparse matrix strategies can yield high precision while cost less CPU time and storage space compared with the concept projection matrix method. This research is based on the algorithm developed in [3]. To further investigate our multistep algorithms' performance and efficiency, we test on two large datasets, LISA and NPL. Through the experiments we see that MSSP algorithm is not only useful for small datasets like MED, CISI and CRAN but also has promising results for large datasets with higher dimensionality.

Although in our study, the performance of MSSP on NPL is not as good as it is on LISA, the advantages of MSSP in saving CPU memory space and computation efficiency are still encouraging. The best precision result of MSSP on NPL can be comparable with that in [7]. In the coming future, the algorithm MSSP is expected to be applied to larger dataset like OHSUMED. And we are seeking ways to parallel the algorithm and make it to speed up the computation at each step.

References

[1] J.Dobsa and B.J. Basic. Concept decomposition by fuzzy k-means algorithm,. *IEEE Web Interlligence Conference (WI2003)*, pp. 684-688, 2003

[2] A. Kontostathis, W.M. Pottenger, and B.D.Davison. Identification of critical values in Latent Semantic Indexing(LSI). *In Foundations of Data Mining and Knowledge Discovery* pp. 333-346, 2005.

[3] C. Shen and M. Unuakhalu. A Class of Multistep Sparse Matrix Strategies for Concept Decomposition Matrix Approximation. *Proceedings of the 2009 ACM symposium on Applied Computing. ACM*, 2009.

[4] E. Chow. A proori sparsity patterns for parallel sparse approximate inverse preconditioners. *SIAM J. Sci. Comput.* , 21(5):1804-1822, 2000.

[5] D.S.Modha and I.S.Dhillon. Concept decompositions for large sparse text data using clustering. *Machine Learning* , 42(1):143-175, 2001.

[6] K. Wang and J. Zhang. MSP:A class of parallel multistep successive sparse approximate inverse preconditioning strategies. *SIAM Journal on Scientific Computing* , Vol. 24, No.4, 1141-1151, 2003.

[7] A. Kontostathis. Essential Dimensions of Latent Semantic Indexing (LSI). *40th Annual Hawaii International Conference on System Sciences*, pp. 73, 2007.

[8] J. Chen and Y. Saad. Divide and Conquer Strategies for Effective Information Retrieval, *Proceedings of the Ninth SIAM International Conference on Data Mining (SDM)*, 2009.

[9] D. Zeimpekis and E. Gallopoulos. TMG: A MATLAB toolbox for generating term document matrices from text collections *Springer*, pp. 187-210, 2006.

[10] P.Hubands, H. Simon, and C. Ding. On the use of singular value decomposition for text retrieval. In e. M. W. Berry, *Computational Information Retrieval* pp. 145-156. Philadelphia: SIAM, 2001.

Cluster Analysis on User Profile Variables for a Digital Media Learning Environment

Arturo Fernandez Espinosa, Meaghen Regts, Jayshiro Tashiro and Miguel Vargas Martin
Faculty of Business and I.T., University of Ontario Institute of Technology, Oshawa, Ontario, Canada

Abstract— *Cluster analysis has been widely used in data mining for its efficiency to group elements based on their characteristics. In this paper we report on the use of cluster analysis to find patterns that help classify a sample of 69 students based on their profiles and their performance on learning activities in a digital media learning environment called IPSims. Our goal is to determine which variables of the user profile have the highest impact in student performance.*

Keywords: Data mining, knowledge discovery, cluster techniques, dimensionality reduction, data management.

1. Introduction

Cluster analysis is an efficient tool used in data mining and knowledge discovery processes. Cluster analysis groups elements according to their characteristics that eventually will allow the researchers to detect and observe specific related to the user profiles and their relationship with the final outcome within IPSims activities.

The main objective of this work is to prove that there is an identifiable subset of variables in the user profile which have the highest impact in students performance within the IPSims activities. Cluster analysis will be used to generate groups of students with identifiable characteristics, expecting that each group reflects noticeable common characteristics inside each student cluster. One of the main problem to address in this research is the selection of a valid subset of variables from the user profile. The original user profile is conformed of 24 variables, that were subsequently reduced to a subset of 13 meaningful variables.

The implementation of K-Means as well as the cluster analysis is another aspect to solve. After generating the clusters of students it is imperative to define the criteria to be used in order to determine which variables are the ones having highest impact in the student performance within IPSims.

IPSims data are stored on a Postgres database that will be processed using Python scripts. The implementation of K-Means will be coded on Python as well. The rest of the paper is organized as follows: Section II review related literature about cluster analysis. IPSims is briefly described in section III. In Section IV the process to detect the high impact variables is explained.Section V presents the conclusions and final notes of this work.

2. Related Work

Cluster analysis has been widely use as a data mining tool. There is innumerable algorithms and fields where cluster analysis can be applied. In most of them the intention is to identify specific groups of objects based on their characteristics. Sometimes, these groups lead to important discoveries such as relationship between variables. De-cheng and Xia [1] make use of cluster analysis on Regional Investment Climate of Mainland China. The intention of this work is to apply cluster techniques in order to divide 31 regions into 5 categories(clusters) based on 43 variables. Their intention is to give advice on how to improve the RIC for the regions in each category. The analysis is done using the SPSS tool and is interesting to see how based in previous studies of regional investment they select the 43 variables that are being analyzed. In the present paper the variables were selected by experts in the educational area and some others were discarded for the lack of variability for each student. Pi et al. [2] proposed novel algorithms for Microarrays using R* trees. The proposed algorithms called KMeans-R and Hierarchy-R. These algorithms are improved versions of K-Means and hierarchical clustering. As mentioned before there is a plenty of clustering techniques, some of them behave better for specific cases, as some may be preferred over others depending on specific scenarios. The algorithms proposed by Pi et al. [2] are suitable in terms of clustering quality for the dataset used in their experiments. In the present work the selection of K-Means was adequate given the simplicity of the algorithm and the reduced number of study objects despite the fact that a bad selection in the number of clusters K will yield poor cluster quality.

Cluster techniques can be applied in a variety of objects Zhou [3] proposed a novel algorithm called Inc-Cluster which showed an improvement in terms of speed against an algorithm called SA-Cluster. Both algorithms are used to cluster graphs based on the vertex properties. It is interesting to see how cluster analysis technique can be combined or can be based in other techniques like bio-inspired heuristics. The present paper does not require to generate novel algorithms or improvements over the traditional ones given the nature and simplicity of the IPSims dataset.

The versatility of cluster analysis make them useful in every kind of environments. A clear example of this versatility is the one shown by Hirano and Tsumoto [4]. They apply cluster analysis in time-series for medical data. The approach is based on multidimensional trajectories making use of a multi-scale comparison technique that compares of the structural similarity of the data.

Yang et al. [5] apply the principles of the bio-inspired heuristic of ant colonies in the analysis of reservoir slopes, in this case he applies the artificial intelligence of the ant optimization algorithms in a cluster technique called fuzzy disturbance ant-cluster algorithm which shows a considerable improvement compared to traditional techniques applied over reservoir slopes. This kind of technique might be ideal for this specific case but the experiments are realized over a single dataset so there is no guarantee that the algorithm shows a better performance compared with the traditional algorithms if applied in a dataset of different nature.

He et al. [6] show the application of cluster analysis on symptoms and signs of traditional chinese medicine in patients with unstable angina. The intention of the author is to provide an exploratory research over symptoms and signs on 815 patients with unstable angina with the intention to provide classification of patients and diagnosis of the unstable angina based on the Traditional Chinese Medicine. We can observe a similarity with our work here, the authors want to generate clusters in order to give diagnostic, while we want to generate groups in order to observe why some students get a given result in their activities within the IPSims.

3. IPSims Description

IPSims is a digital learning environment designed to support and enhance student's learning experience. The use of an interactive environment supported by multimedia materials like videos, medical profiles, local and external information sources used in an effective way in order to facilitate students learning approach. IPSIms has different functionalities to support the instructor and faculty members in the understanding of students academic performance. IPSims tracking algorithm keep track of the activity of the students within IPSIms, as well IPSIms persists a user profile that will be the core of this research. IPSIms first-time users get registered in the system with a user id that will be generated automatically after they fill the user profile and select a password all this presented in a web form generated by FLEX technology [7]. Figure 1 shows the login screen for IPSIms.

The user profile, user id, password and navigation preferences are stored in a secure database that will be the main source of information in the knowledge discovery process used by the authors to discover patterns and relationships between variables.

After the registration process the system will present a menu screen to the users where they can select one of the six simulations scenarios according to the instructors instructions. Each simulation contains three scenarios, a library with web links and scholarly journals, scopes of practice, inter-professional competencies, inter-professional perspectives, case records, case encounter (video), main menu, logout, and bookmark links. Figure 2.

4. User Profile Analysis

The user profile consists of the variables in the login screen and a questionary made by experts in the educational area. Thus, to keep a good ration variable-study cases, we decided to create a subset of variables based on a dimensionality reduction technique. The factor analysis Sheppard [8] was the selected technique to reduce the original dataset.

After applying factor analysis reduction we obtained a Kaiser-Meyer-Olkin measure with a value of 0.541 which means that the feature extraction results will be acceptable. According to the literature values from 0.5-0.7 are mediocre, 0.7-0.8 good, 0.8-09 excellent and greater than 0.9 will reflect superb results in the feature extraction. Our value is in the acceptable level but will retrieve mediocre results in the feature extraction. Once analyzed the correlation matrix, component matrix, as well as eigenvalues and variances the final subset of variables is selected.

The final subset of variables taken from the IPSims database and the survey paper based on the factor analysis are as follows:

- Surf hour per week: Will be shown in the web form as a text box where the user can record any positive number or zero according to the number of hours peer week the user surf on internet. the persistence in the database will be using the same format.
- High school average: This variable is taken from the questionary and the user will write her average from high school based on 5 intervals between 0-100. The variable is persisted as an integer from one to five where one indicates the lowest interval and five the highest one.
- Pursue of post secondary education: This variable indicates if the student is interested on keep with her studies after finishing the university. The variable is taken from the survey paper and is persisted as one if they want to pursue studies after the university and two if they don't want.
- Age: The variable will keep the age of the user, this variable is taken from the survey paper and will be stored in the database as an integer positive number.
- Computer literacy: The variable is stored from the web form and presented to the user as a combo box containing the values: Terrible, Poor, Average, Good, Excellent. The variable is persisted as a number from

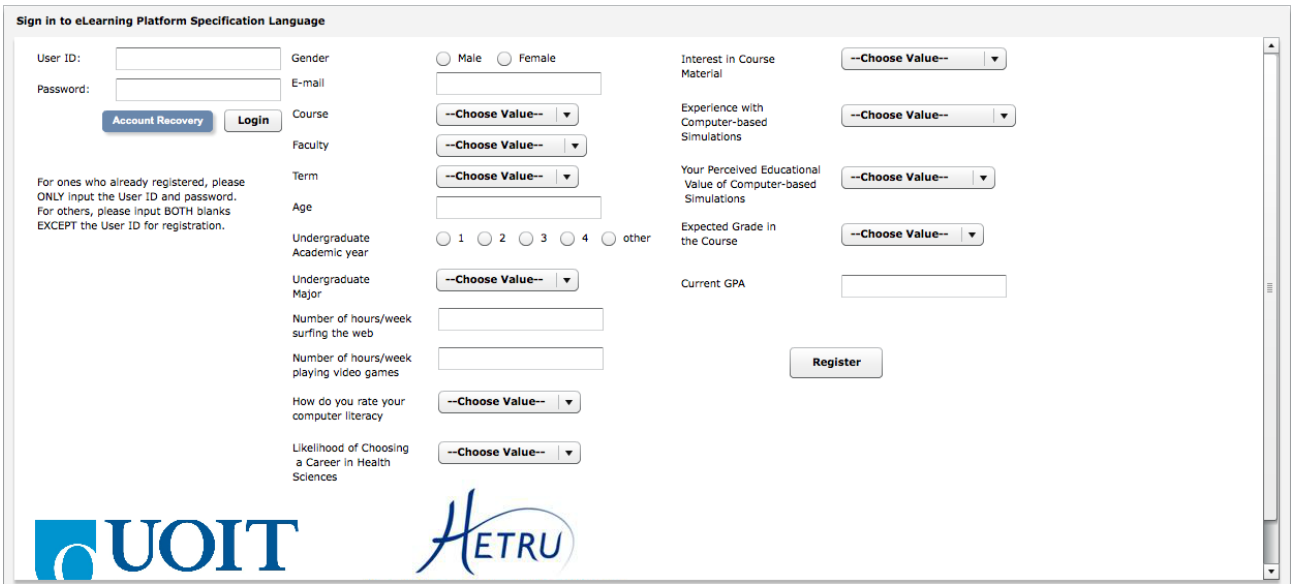


Fig. 1: IPSIms login screen

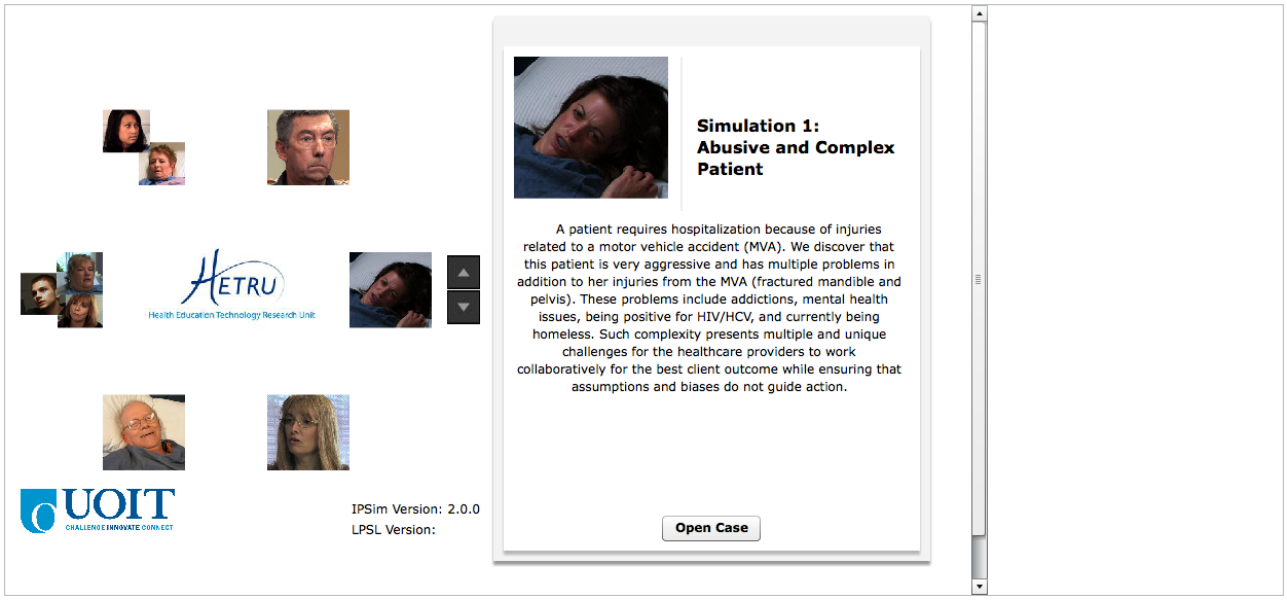


Fig. 2: IPSims main menu

one to five one corresponding to Terrible, two for Poor and so on.

- Likelihood of choosing a career in health sciences: The variables is stored from the web form as well and presented in a combo box displaying the values: Not At All Likely, Not So Likely, Likely, Very Likely. The values are persisted in the database as integers going from one to four starting from one mapping to Not At All Likely.
- College or university before the current university: This variable stores a value of one or two, one if the student took college or university before the current one. and it uses two in the case of have not course any of these mentioned before.
- Gender: This variable stores the gender of the user. Two for women and one for men.

This subset of variables conforms the high dimensional vector used in the cluster analysis.

The web form is completely validated so the users can introduce just valid values so there is no risk of inconsistency in this part of the data acquisition. The questionnaires let some inconsistencies happen such as missing variables for

some students. The use of a well known technique to dismiss outliers was used, the mean for each variable was calculated and 5% of the top and bottom values were substituted with the calculated mean.

5. Data Collection

The strategy to collect the data was planned to be within a couple of courses in health sciences. The students from the class had an information session to show them how to use the system, as well a paper questionary was given to them. This questionary were related with a user profile and usability questions as well another questionary was handled in order to evaluate their performance within the activity.

After prepare the students to use the system. They filled the paper questionary with the questions related to their user profile. Then they proceed to generate their user profile Figure 1 to access the system in their computers. They were required to accomplish with the activities that conformed one of the scenarios presented on IPSims and they have to answer a paper questionary to evaluate their performance within the activity.

69 students were recruited within the two different courses of health sciences. The information was treated and cleaned from inconsistencies and outliers. All the variables from the paper questionary were persisted into a database that eventually was merged with the existent information in the system related with the user profile generated in the main screen of IPSims Figure 1.

6. Cluster Analysis

The clusters were generated by the well-known K-Means algorithm. Five clusters were generated using as identifier for each study object the user ID. After the five clusters were formed we proceeded to determine the means of each variable by cluster.

The variable grade is stored in the IPSims database, this variable stores the final grade for the activity corresponding to each user. This variable will be determinant to our analysis and it has values that vary from 0 to 18 being 18 the highest grade a student can get in the activity and 0 the lowest one. To test the hypothesis that there exists a set of profile variables that have a direct impact in the final outcome (grade) of a student in IPSims we adopt a straight forward strategy. After obtaining the output of K-means we calculate the mean of each variable by cluster as well the mean grade for each cluster. Then the clusters are organized from lowest to highest grade.

In order to recognize the high impact variables the authors are expecting to see a direct correlation between variable means and the mean of the grade for each cluster. This analysis yield some interesting results. This results are observed in the plots of the calculated mean of some given variable versus The calculated mean grade for each cluster.

Figure 3 and Figure 4 are some of the resulting graphs from the described processes mentioned before.

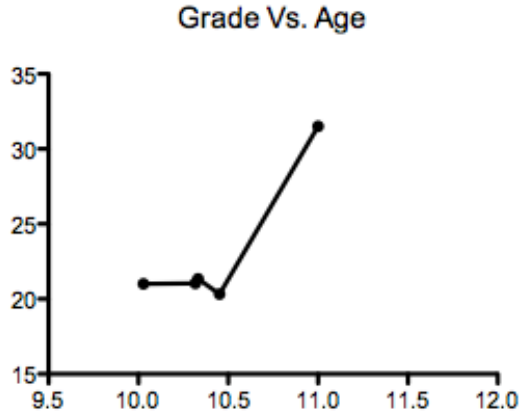


Fig. 3: Grade Vs. Age plot

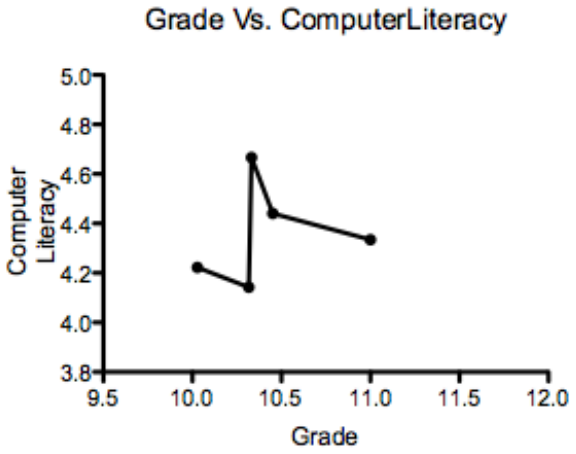


Fig. 4: Grade Vs. Computer Literacy plot

After analyzing and observing the generated plots we discovered some interesting patterns that could support our assumptions of a descriptive subset of elements that has the highest impact in the final grade activity. For example Figure 5 show us that the cluster with most elements that reflect plans to pursue a post education studies are the ones with highest grade in the activity as well the cluster that reflects the lower mean in post education studies pursue is the cluster with the lowest average grade.

Figure 6 is another plot that throws interesting results. Is observable that the cluster with the highest average age is the one with the best results in the activity. These kind of graphs give us a positive indicator that it is possible to find

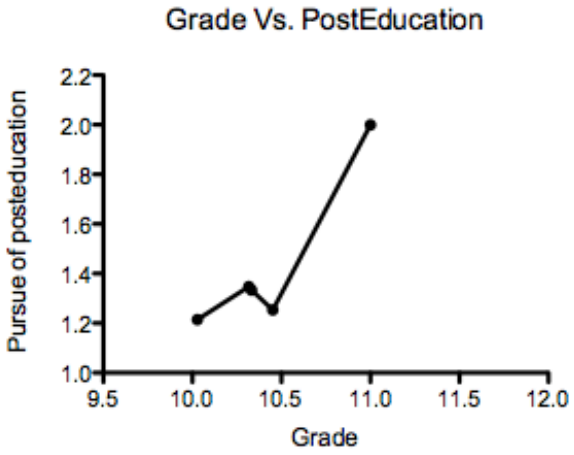


Fig. 5: Grade Vs. Pursue of Post Education Studies plot

a meaningful smaller subset of variables that will have the highest impact in the student activity final grade.

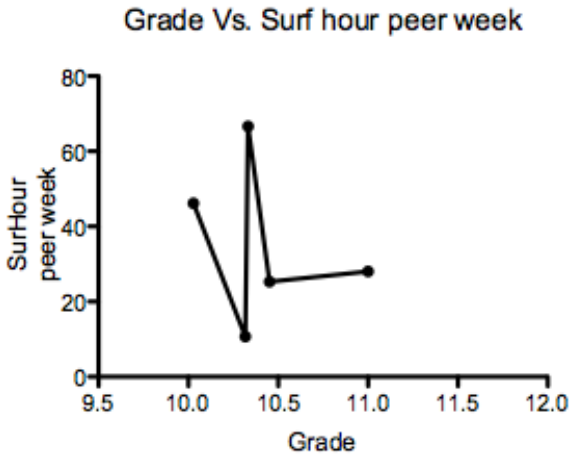


Fig. 6: Surf hour peer week Vs. Grade plot

We follow the same logic to analyze the other variables within the clusters in order to find which variables could be the ones that are having a real impact in the final grade of the activity within IPSims.

The generation of 3D graphs allow the researches to locate some interesting regions where a correlation between tuples of variables and the grade can be observed. Figure 7 and Figure 8 shown some interesting arrangements where we can observe some defined patterns or behaviours between groups of variables in function of the grade.

7. Conclusions

The intention of this study was to test that there is a subset of variables in the user profile that has a direct impact in

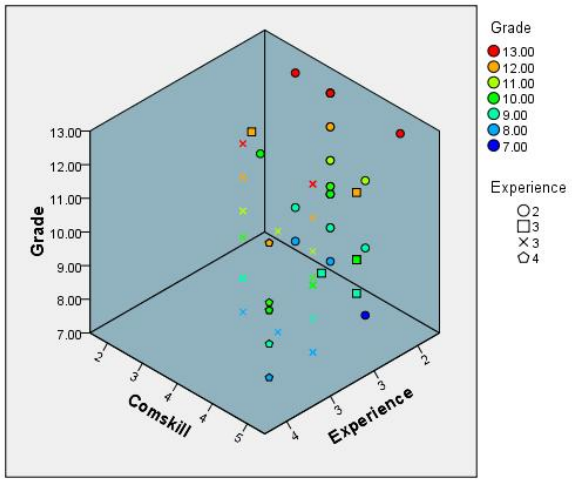


Fig. 7: Grade, age and high school average plot

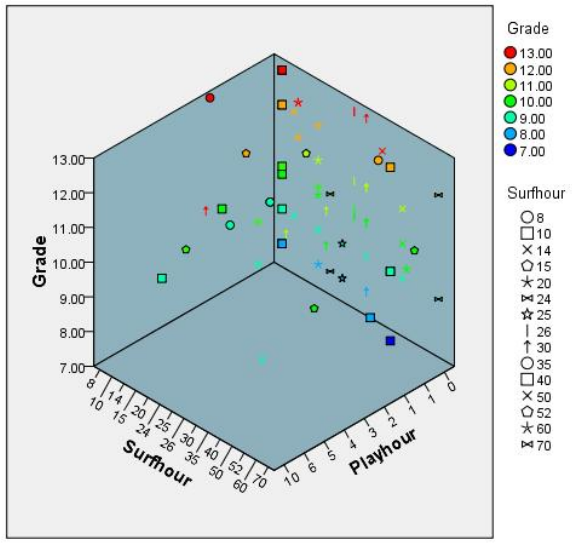


Fig. 8: Grade gender and surf hour peer week plot

the final grade of the student within the digital learning environment called IPSims. After the factor analysis for dimensionality reduction we can observe that the selected subset still reflects interesting patterns that allow us to think that from that selected subset there is still a reduced one conformed by variables like age, pursuing of post education and some others that has the highest impact in the student final result within the activity.

The variables selected after the factor analysis were: surf hour peer week, high school average, intention to pursue post education, college or university before the actual institution, age, gender, computer based simulations literacy

and likelihood for course material. From these variables and the proposed analysis the ones showing the most interesting patterns were: high school average, intention to pursue post education, age and study of college or university before the actual one.

Using the 3D scatter plots we observed some interesting patterns that let us know that there is some interesting codependencies between variable tuples and the final grade in the activities. For example was noticeable that older students with precedent studies in a college or another university has better results in the activities.

From here the authors are establishing the starting point to propose a prediction model based in the user profile.

Unfortunately the number of study cases is low counting just with 69, and the variability in the marks for the activity is low too, so the variable grade is reflecting means for each cluster that are really close having values just from 7 the lowest one and 13 the highest one. There is predomination of nines and tens. There is interesting patterns showed by the constructed plots after applying the k-means algorithm but in the future a bigger variability within the user profile variables is expected given the increment in the study cases. This variability will allow to accomplish with the discovery of different patterns and the construction of a bigger number of clusters.

The project expects to receive more study cases in next semesters in order to make the variability of the variables wider. That will allow the researchers to discover more patterns that will support the results obtained on this first phase of the research.

The observed patterns suggest the authors that some measures could be taken for the instructors in charge of the courses in order to improve the performance of the students within the IPSims which will have an impact too in their apprenticeship of the subject and will improve their experience in a digital learning environment that prepare them for real life situations.

The project has being planned to generate more knowledge with the addition of more study cases in the next scholar year. With the use of more complex tools and information visualization techniques IPSims will become a valuable source of information that will support the instructors and will work as a research tool in the education and computational science areas.

Acknowledgment

This work was supported in part by the Social Sciences and Humanities Research Council of Canada.

References

[1] FAN De-Cheng and PAN Xia, “Appraisal and Cluster Analysis on Regional Investment Climate of Mainland China”

[2] Jiaxiong Pi, Yong shi and Zhengxin Chen, "Similarity and cluster analysis algorithms for microarrays using R* trees," IEEE Computational Systems Bioinformatics Conference Workshops, IEEE. 2005

[3] Yang Zhou, Hong Cheng, and Jeffrey Xu Yu, "Clustering Large Attributed Graphs: An Efficient Incremental Approach," IEEE International Conference on Data Mining IEEE. 2010

[4] Shoji Hirano, and Shusaku Tsumoto, *Sixth International Conference on Data Mining*, 2006.

[5] Yang Yang, Zhao Eerfeng, Fang Chunhui, Wang Yachao and Xu Baosong, “Analysis of Reservoir Slopes based on Fuzzy Disturbance Ant-cluster Algorithm,” *IEEE 2011*

[6] Wingyong He, Jie Wang, Yunling Zhang, Yanli Tang and Yue Zhang, “Cluster Analysis on Symptoms and Signs of Traditional Chinese Medicine in 815 Patients with Unstable Angina,” *IEEE Sixth International Conference on Fuzzy and Knowledge Discovery*, pp. 435–439, 2009.

[7] Sheppard A. G., "The sequence of factor analysis and cluster analysis: Differences in segmentation and dimensionality through the use of raw and factor scores", *Tourism Analysis*, 1 pp. 49-57.

Tracking Sentiment Analysis through Twitter

Thomas Carpenter and Thomas Way
Applied Computing Technology Laboratory
Department of Computing Sciences
Villanova University, Villanova PA 19085
thomas.way@villanova.edu

Abstract – *Social media continues to gain increased presence and importance in society. Public and private opinions about a wide variety of subjects are expressed and spread continually via numerous social media, with Twitter being among the most timely. The ability to quantify and evaluate society's perceptions is increasingly critical to the success in the marketplace. Sentiment analysis is an approach that can be used to computationally measure perceptions regarding topics based on selected, textual source material. This paper reports on the design of a sentiment analysis tracking algorithm and its implementation in the form of a web-based application that can quantify sentiment contained in Twitter feeds and track the sentiment as it changes over time. Results are provided that evaluate the tool for use in performing tracking of sentiment analysis as perceptions change over time.*

Keywords: Sentiment Analysis, Twitter, Opinion Mining, Sentiment Tracking.

1 Introduction

People have always been interested in what others think. It is human nature to seek the opinions of others, and with the availability of vast online resources for expressing opinion, it is increasingly feasible to automate the process of discerning widespread opinion or sentiment on any given topic, idea, product or person. With the dramatic growth of social media in the past 10 years, an aggregated study of social media statistics that may be beneficial to business found that over 80% of active online users make use at least one form of social networking and nearly 23% of time spent by those users online is spent on social networking sites [6]. More specifically, the study found that 51% of Facebook users and 64% of Twitter users are more likely to buy from the product brands they follow.

It is well known that social media sites such as Twitter and Facebook are frequently used to express, or post, opinions about a particular topic of interest to the poster. These opinions can be a rich source of feedback to the marketplace if they can be gathered and analyzed in a timely and meaningful way. The field of sentiment analysis, which includes opinion mining, provides a variety of approaches that can attempt to manage and make sense of this large and widely distributed resource of opinion. [9]

While sentiment analysis of one form or another has been investigated as early as 1979 [1], research in the area has increased dramatically since 2001. Since that year, with concurrent improvements to machine learning and NLP techniques, the widespread use of the Internet, the ease of publishing material online, and the recognition that understanding trends in public opinion expressed online can be a valuable source of information to business and researchers [9].

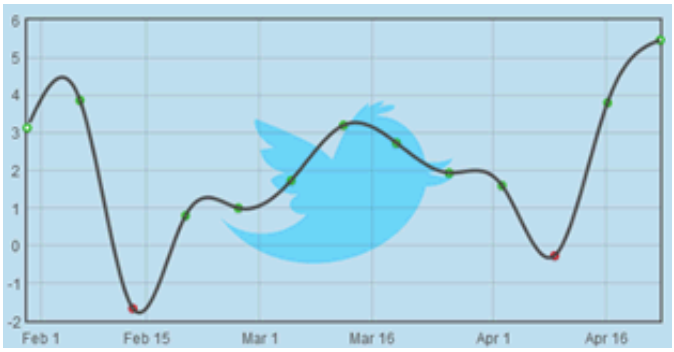


Figure 1. Example sentiment tracking graph.

In this paper, the design of an algorithm that performs sentiment analysis over time using data from Twitter is introduced and a web-based application called the Villanova Analytical Sentiment Tracker (VAST) that implements the algorithm and generates sentiment tracking graphs (Figure 1) is described. Results of preliminary tests of this sentiment tracking approach are discussed, and plans for future development of an online sentiment analysis research tool are presented.

2 Quantifying Sentiment

The challenge of performing sentiment analysis tracking on a large set of source material, such as that available from online social media and over a given period of time, can be broken down into two complementary tasks: Sentiment analysis and sentiment tracking.

2.1 Sentiment Analysis

Sentiment analysis is an active area of research that makes use of natural language processing techniques to quantify an expressed opinion or sentiment within a selection of text. [8] Common uses of sentiment analysis include managing and analyzing review-related online sources, computationally verifying partial content as it relates to overall content in a source material, assessing marketplace reactions to business practices and products, measuring public perceptions of political figures and celebrities, and a wide range of diverse uses [9]. There are a wide variety of algorithms used for performing sentiment analysis, from simple word occurrence counting to machine learning [9], with the selection of one or more approaches typically dependent on characteristics of the material being analyzed [9].

The advent and widespread use of social media, with its emphasis on expression of individual opinion, has provided a rich source of material for large-scale sentiment analysis [2]. Social media has increasingly become a tool for spreading interest in a product, person or idea, and for general promotional use within the online marketplace [5]. If there are negative blog posts or tweets about a product or business, there is a good chance that the business will notice a decrease in their sales. Key among the reasons why this is the case is because those who tend to use online resources frequently to gather and post opinions tend to place significant weight in the information they gather [3]. Thus, it is in the interest of the online marketplace in all its forms to pay attention to the opinions expressed online via social media.

Although sentiment analysis is prone to the same difficulties that general natural language understanding approaches are, the problems are well-understood and solutions to manage these challenges and provide statistically meaningful results exist [2,9]. The use of Twitter as a time-stamped data source has been identified as a realistic and substantive resource [7], as microblogging in general has found a place in online brand marketing [4].

2.2 Sentiment Tracking

Periodically recording the results of Sentiment Analysis over time provides a mechanism for assessing opinions about a topic of interest and quantifying the impact of outside forces, such as current events, popular trends and marketplace competition. Sentiment Tracking gathers the output of Sentiment Analysis at a desired time granularity and over a desired time duration to provide time-based sentiment data for a desired topic.

This longitudinal data can then be analyzed for simple trends or can be compared against other external events. For example, performing daily sentiment analysis on a prominent politician over a period of one year can be juxtaposed with the activities of that politician. This juxtaposition can be analyzed to attempt to quantify the impact of those activities or other current events on the public perceptions expressed via social media.

2.3 Applicability of Sentiment Tracking

An approach such as sentiment analysis and tracking can be appealing to a wide variety of audiences. The general public may be interested in checking what people are saying about their favorite pop star. Celebrities, and the people who hire them and pay them, can see what a large group of people think of them. Politicians may want to know what people are saying about them during an election. Sports analysts want to know what percentage of people are supporting a certain team during a game, which is an opinion that can fluctuate even while the game is happening. Businesses may desire to know if people are upset with them during the latest fall in stock prices. These factors and many others provide strong motivation for the implementation of a flexible, easy to use, sentiment tracking tool.

3 Sentiment Tracking Tool Design

The tool we have designed to perform sentiment tracking using Twitter is called the Villanova Analytical Sentiment Tracker (VAST) tool. The VAST tool is composed of two parts: tracking and evaluating. The tracking portion of the tool provides the ability to quantify the trend of “positive” or “negative” tweets over a specified period of time. The evaluation part of the tool involves the performance of sentiment analysis to attempt to discern a negative or positive opinion in each tweet.

3.1 Motivation for Tool Design

The purpose of developing the VAST tool was to create the ability to track any specified topic on Twitter over any given period of time within the lifespan of Twitter. While there are other tools easily found with a Google search that perform Twitter analysis in real-time, that was not the intent of the design of the VAST tool. Rather, we have aimed to develop a platform for research and further development of sentiment analysis algorithms and sentiment tracking approaches that would be useful to researchers in many fields.

3.2 Description

The VAST tool allows a user to enter a search term, a start and end date, an interval period of time, the number of tweets for each period, and a threshold accuracy level (Figure 2). The start and end dates refer to the overall time period to track the search term. The interval period refers to how often the user wishes to return results within the two dates (daily, every 2 days, weekly, bimonthly, monthly, or yearly). The number of tweets variable refers to the number of tweets, positive or negative (neutral should not be returned, if possible), that the user wishes to return for each time interval.

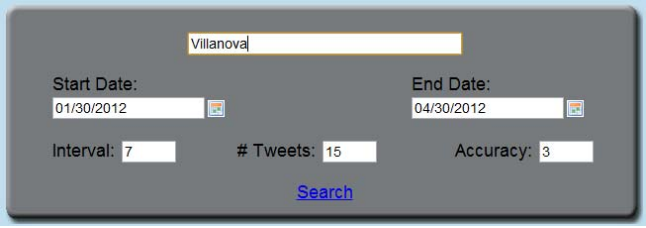


Figure 2. VAST tool control panel interface.

For example, if a user inputs a weekly time interval with 15 tweets per week, the system will return the 15 tweets, classified as either positive or negative. These tweets will be what are determined as the most popular tweets (most frequently re-tweeted or other measurement of popularity) and with the highest sentiment values, for each week in the overall date interval.

Finally, the accuracy parameter refers to the threshold the system uses for determining if a tweet is positive, negative, or neutral. If the accuracy is set to 3, for instance, then the system will only return tweets that have a sentiment value outside the range from -3 to +3. The higher the accuracy is set, the longer the system will take to find the desired number of tweets with that accuracy level. In preliminary test runs, it was determined that an accuracy level of 3 produces acceptable results in terms of speed, quantity of output and quality of output (as determined by qualitative, manual analysis).

Because some sentiment tracking results can take quite some time to produce (minutes to hours), the VAST tool is designed to display results dynamically. As the result of analysis is completed on each time interval, the results are immediately displayed. The user of the VAST tool thus is able to visualize and evaluate the results in an ongoing fashion, enabling interactive analysis as more results are displayed or early termination of the results appeared incorrect in some way.

3.3 Classification Approach

Through experimentation with a number of sentiment analysis classification techniques, it became clear that in order for the VAST tool to perform tracking in the way we proposed it was necessary to develop an algorithm specifically for the needs of the tracker (Figure 3, following page). The more common language classification techniques described in the literature and used for sentiment analysis were either too inefficient or did not provide exception handling for situations where our approach required customized analysis. The resulting algorithm makes use of five word lists: ExtraWordList, MainWordList, NegationList, AdvList, IntensifyList.

The ExtraWordList is the first step of the algorithm. It holds a combination of two types of words or phrases. Its first function is to hold special cases involving the search term inputted to check for in the tweets. For example: [SearchTerm + " beat"] should actually be a positive sentiment because the tweet is most likely referring to a team or politician winning some sort of competition.

Table 1. Sample phrases and their weighted score from ExtraWordList.

Phrase	Score	Phrase	Score
Go + SearchTerm	+4	Don't care	-2
SearchTerm + Beat	+2	No thanks	-5
Beat + SearchTerm	+2	Pretty much	+1
Killing	-3	Does stuff right	+2

```
FOR every Tweet in the List returned from API:

   FOR each word in the Tweet:
    IF the word is in the NegationList
      NegNum = -1
    ELSE IF the word is in the AdvList:
      advNum = 1
    ELSE:
      IF the word is in the MainWordList:
        IF Word value is Positive:
          PosCount = PosCount *negNum+advNum
        ELSE:
          NegCount = NegCount *negNum-advNum
        IF NegNum == -1:
          Append "NOT "+ Word to SentimentWordsList
        ELSE IF AdvNum == 1:
          Append "Adv  "+ Word to SentimentWordsList
        ELSE:
          Append Word to SentimentWordsList
      #Reset Variables
      NegNum = 1
      AdvNum = 0

    FOR word in IntensifyList:
      IF word is in Tweet:
        IF tweet is Positive
          PosCount = PosCount + 1
        IF tweet is Negative
          NegCount = NegCount - 1

   # Add as Positive tweet
   if posCount+negCount  >  accuracy:
          positive ++
           Append Tweet and Tweet Data to Tweet Array
   # Add as Negative tweet
   if posCount+negCount  <  -accuracy:
           negative ++
           Append Tweet and Tweet Data to Tweet Array
   # Add as Neutral tweet to separate Neutral Array
   else:
            neutral ++
            Append Tweet and Tweet Data to Neutral Array List

If there aren't Enough Positive and Negative Tweets to return
     Append Highest Sentiment Valued NeutralTweets to Tweets Array
If there are too many tweets in Tweets Array
     Return the first (numTweets) tweets from Tweets Array
```

Figure 3. Pseudo-code algorithm for Sentiment Tracking in VAST tool.

However, when applying the MainWordList (Table 2) later, this tweet may be considered negative because "beat" is considered a negative word. By adding phrases like this, it is possible to even out or offset the value of the tweet to portray the true sentiment to the tweet. The second type of entry in the ExtraWordList is similar to the first, but without the SearchTerm. The MainWordList does not deal with phrases, only single words.

Table 2. Sample phrases and scores from MainWordList.

Phrase	Score	Phrase	Score
Adorable	+3	Sluggish	-2
Adore	+3	Slut	-5
Kill	-3	Smart	+1
Killing	-3	Smarter	+2
Miracle	+3	Smile	+2
Misbehave	-2	Smiled	+2

The ExtraWordList is used to add in phrases that may invert the results of a negative word to actually be positive and vice versa. For example the phrase "pretty much" will return a small positive value since the word pretty is considered positive. Since this really should just be neutral, the phrase "pretty much" was added to the extra words list with the value of -1 in order to offset the +1 value that will be associated with the word "pretty" later on.

The reason that the algorithm is able to use phrases in the ExtraWordList and not the MainWordList is due to the method of finding the words in the tweet. Since ExtraWordList is very small, the system can efficiently loop through each word/phrase in the list and check if the whole phrase is in the tweet. If the system were to do this on the MainWordList it would have to run through the MainWordList of thousands of entries many times, which would significantly hurt performance. Instead, the algorithm loops through each word in the tweet (a relative small value of N), checking for occurances found in the MainWordList using a much more efficient hashed dictionary lookup.

Tweets can include a wide variety of symbols, letter, numbers, and punctuation. Therefore, before performing analysis using the MainWordList, it was necessary to remove all characters that were neither alphabetical nor numerical.

Once the non-alphanumeric characters are stripped out, the algorithm iterates through each word in the tweet and quantifies its sentiment. This sentiment quantification loop make use of three of the lists. First, the algorithm checks to determine if the word is a negation word found in the NegationList (Table 3). If it is, the algorithm will negate the NegNum variable used in its sentiment score calculation so that if the next word in the tweet is positive it will negate it, and vice versa for a negative word value.

Table 3. Sample from NegationList.

Not
Doesn't
Wouldn't
Cannot
Can't

Second, the algorithm checks to determine if the word is in the AdvList (Table 4) and if found the algorithm modifies the AdvNum variable used in its sentiment score calculation from 0 to 1. The effect of this adjustment is that if the next word is found in the MainWordList, its score will be weighted as a little more positive or negative.

Table 4. Sample from AdvList.

Really
Very
Extremely
Clearly
Certainly

Finally the algorithm checks if the word is in the MainWordList and if it is, it performs the follow action:

```
PosCount = PosCount +
   (wordValue * negNum) + advNum
```

Or,

```
NegCount = NegCount +
   (wordValue * negNum) - advNum
```

If there was not a negation word or adverb before the word in question then the extra variables will have no bearing on the result since it would read something like:

```
PosCount = PosCount +
   (wordValue * 1) + 0
```

However, if, for example, there is a negation word it would negate the positive value and read:

```
NegCount = NegCount +
   (wordValue * -1) + 0
```

The final List that is used in the algorithm is the IntensifyList (Table 5). This list consists of words that portray strong emotion, but could be either negative or positive depending on the rest of the tweet.

Table 5. Sample from IntensifyList.

Wow
OMG
Holy Sh#t
Unbelievable
Insane

A commonly occurring example of the use of an intensifier is the word "wow." For example, the following are positive and negative tweets, each using the word "wow" as an intensifier.

An example of a positive tweet:

Wow! I can't believe we won the Super bowl! Go Giants!

An example of a negative tweet:

Wow! You're crazy if you use Google Drive!

These words should still be applied as they certainly declare sentiment, but it is likely best to wait to apply them until the overall sentiment of the tweet is determined, as they serve primarily as amplifiers or intensifiers. Once the algorithm has completed analysis to determine the overall sentiment of an entire tweet, it checks to see if the IntensifyList words are in the tweet. If they are, the algorithm adds +1 to the value if it is positive or -1 to the value if it is negative. In this way, the score of the word corresponds to the sentiment of the rest of the tweet. Often, the use of an intensifier can tip the score to clearly positive or negative for tweets that are borderline neutral.

At this point, sentiment analysis of the tweet is complete and the algorithm performs a final check to determine if the tweet is within the threshold accuracy level. If the sum of the PosCount and NegCount values is greater than the specified threshold, then the tweet is quantified as a positive tweet. If the sum of PosCount NegCount value is less than the threshold, then the tweet is quantified as a negative tweet. If the sum of PosCount and NegCount do not clearly make the tweet positive or negative, the tweet is stored in a list of neutral tweets, which is only used in the event that there are too few positive and negative tweets to meet the desired count for the current interval.

4 Results

Based on preliminary results (see example result sets in Figures 4-7) of running approximately 50 terms through the VAST tool, the accuracy of the sentiment analysis portion of the algorithm for correctly quantifying the sentiment of each tweet is in the range of 65-85%. This accuracy level was surprisingly good, given that our expectations were that accuracy would be low due to the difficulty with interpreting the small size of tweets and the well-known inaccuracies inherent in many sentiment analysis approaches in general. For example, the use of sarcasm in a tweet can turn a seemingly positive sentiment into a clearly negative one:

I love you, Obama! Please take more of my Money!

For a sentiment analysis system to analyze these words, even applying more sophisticated techniques such as using N-grams or machine learning, it will almost always come out positive. Humans still have the edge in sentiment analysis, it seems, at least where sarcasm is concerned.

Result Set 1
Search Term: **Villanova**
Dates: 03/01/2012 – 04/30/2012
Weekly Interval
15 Tweets per interval
Accuracy of **3**
Correct: **76**
Incorrect: **22**
Unsure: **8**
Irrelevant: **9**
76/115 =
66% correct

Figure 4. Result Set 1 from VAST tool.

Result Set 2
Search Term: **Villanova**
Dates: 03/01/2012 – 04/30/2012
Weekly Interval
15 Tweets per interval
Accuracy of **4**
Correct: **97**
Incorrect: **16**
Unsure: **5**
Irrelevant: **8**
97/127 = **76.4% correct**

Figure 5. Result Set 2 from VAST tool.

Result Sets 1 and 2 had identical search options except for the accuracy. Result Set 1 had an accuracy of 3 and returned only a 66% success rate, while example 2 had an accuracy of 4 and returned a 76% success rate. This may be an indicator that the accuracy threshold is a key element to the accuracy of the algorithm, which was the intent of the accuracy threshold. While the use of this threshold slows the algorithm slightly, the improved accuracy appears to be worth the extra time.

Result Set 3
Search Term: Google Drive
Dates: 04/24/2012 – 05/03/2012
Daily Interval
15 Tweets per interval
Accuracy of 3
Correct: 59
Incorrect: 6
Unsure: 8
Irrelevant: 5
59/78 = 75% correct

Figure 6. Result Set 3 from VAST tool.

Result Set 4
Search Term: 'Somebody that I used to know'
Dates: 01/30/2012 – 04/03/2012
2 week Interval
15 Tweets per interval
Accuracy of 4
Correct: 87
Incorrect: 6
Unsure: 6
Irrelevant: 0
Neutral: 8
87/107 = 81.3% Correct

Figure 7. Result Set 4 from VAST tool.

The data from Result Sets 3 and 4 (Figure 5) demonstrate that the overall algorithm performs accurate-enough sentiment analysis given its simplicity. Further improvements are planned to improve accuracy and efficiency of the algorithm.

5 Conclusions & Future Work

Sentiment Analysis and Sentiment Tracking holds significant promise for analyzing the perceptions of people who use social media. With the pervasiveness of social media, such as Twitter, Facebook, and the growing popularity of blogging in general, these forms of analysis are viable techniques for quantifying the sentiment of the public at large on a given topic. A tool that can efficiently and accurately quantify human emotion is a challenge to implement. The algorithm that is implemented in the VAST tool reported in this paper attempts to perform sentiment tracking, and has achieved reasonably good accuracy in preliminary experiments.

Research in the area of sentiment analysis is active, and we plan to continue to develop the VAST tool and provide it as a research platform. The VAST tool and algorithm it implements can be made more efficient, and we plan to investigate other sentiment algorithms that are especially accurate for shorter source texts such as tweets.

Other planned extensions of the VAST tool include incorporating GeoLocation data to segment sentiment based on geographic region, data export to provide a means for analysis of sentiment tracking results with other software, and making use of Google search APIs to broaden the scope of sentiment tracking beyond Twitter to the Internet at large. Finally, we plan to collaborate with colleagues in other disciplines. For example, as this is an election year, we intend to coordinate sentiment tracking with colleagues in Political Science to determine the post-mortem, and possibly even predictive, applications of sentiment tracking in an election.

6 References

[1] Jamie Carbonell. Subjective Understanding: Computer Models of Belief Systems. PhD thesis, Yale, 1979.

[2] Namrata Godbole, Manjunath Srinivasaiah, and Steven Skiena. Large-scale sentiment analysis for news and blogs. ICWSM, Boulder, Colorado, 2007.

[3] Paul Hitlin and Lee Rainie. The use of online reputation and rating systems. Pew Internet & American Life Project Memo, October 2004.

[4] Bernard J. Jansen, Mimi Zhang, Kate Sobel and Abdur Chowdury. Micro-blogging as online word of mouth branding. In CHI EA '09: Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems, New York, 2009.

[5] Lillian Lee. "I'm sorry Dave, I'm afraid I can't do that": Linguistics, statistics, and natural language processing circda 2001. In Comptuer Science: Reflections on the Field, Reflections from the Field, pages 111-118. The National Academies Press, 2004.

[6] Phil Mershon. Social Media Stats for Small Businesses. Social Media Examiner, Nov. 8, 2011. Available at: http://www.socialmediaexaminer.com/26-promising-social-media-stats-for-small-businesses.

[7] Alexander Pak and Patrick Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Twitter as a corpus for sentiment analysis and opinion mining. In Proceedings of the seventh international conference on Language Resources and Evaluation (LREC), May, 2010.

[8] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79-86, 2002.

[9] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2):1-135, 2008.

The Untold Story: Collecting Tweets from Twitter

Ahmed Nagy
IMT Institute for Advanced Studies Lucca
Italy
ahmed.nagy@imtlucca.it

Natalie Bennett
Carnegie Mellon University
NASA Mountain View
natalie.bennett@sv.cmu.edu

Jeannie Stamberger
Carnegie Mellon University
NASA Mountain View
jeannie.stamberger@gmail.com

ABSTRACT

The paper examines methodological pitfalls for rigorously collecting Tweets related to disasters, including those not otherwise described in Twitter developer forums. Twitter has been an excellent resource for information to analyze informatics and social behaviors in disasters; however, there is little work describing and comparing methodologies for collecting Tweets. (e.g., Twitter search box, Twitter API, and 3^{rd} party archiving tools). Currently, cleaned datasets of Tweets to help standardize research have been promised by NIST and Library of Congress but are not yet available to researchers. Lack of clear well-described methodologies for collecting Tweets creates difficulties in comparing results among studies. This paper serves as a guide for new researchers interested in collecting Tweets and a reference for Twitter data collection methods, to assist in developing standards and robustness across Twitter research. We are giving the code to the community under the open source license.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval—*social media*

Keywords

Crisis management, Disaster response, Emergency management, Data collection, Twitter, Short messages

1. INTRODUCTION

Social media has become a popular method that citizens use to express their opinions and share information. Twitter, Facebook and YouTube are rich channels for sharing information. Crises events are not an exception; there is a lot of information that is shared and can be collected from social media. We focus on Twitter as a channel where data is disseminated and shared. Recently, researchers have begun studying the use of Twitter under crises and disaster situations. These include hurricanes [3], flooding and wildfires [8], and Haiti earthquake [6]. More examples of using Twitter under crises scenarios are included in [1, 2, 5, 8]. However, most of the studies supply limited information about the data collection phase. This paper is organized as follows: Section 2 presents relevant prior work and section 3 presents our detailed methodology. Section 4 highlights the algorithm we developed and explains our approach with analysis for the challenges faced. Section 5 discusses design decisions. Section 7 closes with the conclusion and ideas for further research.

2. RELEVANT METHODS FOR DATA COLLECTION

Data collection is a crucial stage in carrying rigorous research. There are several ways to collect twitter data. At the time of carrying out this research Twitter was not selling data to individuals. In November 2011, Twitter granted the right of selling tweets to Gnip. Buying 10 % of the firehose costs $5,000 in real time while halfhose costs $360,000 per year, which might be expensive for several researchers especially with the new budget cuts in universities [1]. Another option is getting tweet archives from the National Institute of Standards and Technology-NIST. NIST provides a sample of the tweets, around $16M$ collected in approximately one month. NIST is a good option if there is no need for real time collection. The other alternatives are to build your own crawler or to use some of the APIs that are built to collect the data. The main disadvantage of using ready made APIs appears when Twitter changes the APIs. This renders the API's obsolete. In fact this happened very recently leaving us with the only choice of building our own technique to collect the tweets. We decided to share our experience with the researchers who would like to avoid pitfalls and long development time. We are sharing with the community the lessons we learnt and the code we developed.

3. DETAILED METHODOLOGY

We divide the tweets into three types:

1. Authoritative Tweets - originally authored tweets. After the authoritative tweets are collected they are processed as illustrated in section 4.

2. Re-tweets - RTs by the authoritative handle. RTs by the authoritative handle are stored in the database but

[1] `http://gnip.com/`

not processed to identify discussion (current methods to find discussion include comments and RTs). We have a RT count, which tells how many times it was re-tweeted by everyone up to a 100 times (maximum number reported for tweets is 100 more than that is reported as 100+). We consider a tweet to be popular if it was re-tweeted more than 20 times. This is an adjustable parameter according to the problem addressed.

3. Conversations- Replies that are conversations between a handle and a public user

We recorded information to help us find the original Tweet e.g., the Tweet is in reply to another Tweet, which can help track back the conversations. After Tweet collection all the above types of Tweets are stored in the database. The processing steps below apply only to the originally authored tweets. Following a workflow that will explain the steps used to collect the historic tweets. We highlight whether the step belongs to Historic Tweet collection or Regular Tweet Collection.

3.1 Authoritative Handles Processing

The steps that we follow to handle and process the authoritative tweets can be summarized as follows

1. Pull a table of the authoritative Twitter Handle metadata; this is done only once at the initialization phase.
2. Find authoritative Tweets: During each cycle, Tweets are pulled from official handles using the REST API, these are called Authoritative Tweets.
3. Find Tweets which discuss Authoritative Tweets.

Further analysis revealed that entry of a verbatim Tweet into the Twitter Search API does not necessarily capture all Tweets which contain the verbatim language. Therefore, we developed methods and algorithms to increase the recall level of the results.

3.2 Preparing Query Using the Search API

We use the SEARCH API to find Tweets which discuss Authoritative Tweets. To get the search query, the following are removed from the Authoritative Tweet:

1. compressed links eg `http:\bit.jy.fdsjh` or `goo.gl/fb/fwiQl`
2. hash tags eg #cmusv
3. mentions eg @google
4. non alphanumeric characters eg -help me or ....help me or help me..... (this means remove - or .....) Applying the above techniques generates a very broad list of Tweets including those which are not referencing the Authoritative Tweet. The use of the SEARCH API sometimes does not return all the results that you would obtain if you used the Search field on Twitter.com. We check rate limits after every pull. Twitter generates error messages when the service is down or when the rate is limited etc.
5. Identify the discussion Tweets. A discussion Tweet is identified by an authoritative Tweet that was changed. We consider changes as additions to the tweet. In other words, after getting the set of longest common strings, we perform a set difference between the discussions Tweet and the original tweet. The goal in this step is to detect the similarity and the differences between an original Tweet and a modified one. We developed an algorithm which is a variation of the ''difference" algorithm between two sentences, built on the longest common sequence [7]. For more details on the Difference algorithm check [7]. The algorithm is used to identify overlapping parts between two tweets. The following section explains how the algorithm works.
6. Conduct sentiment analysis on the added text discussing the original Tweet. The Historic Data collection will collect the prior 7 days of data, which allows pulling up to 3200 Tweets. After obtaining the Tweets, we regularly ping Twitter for new Tweets.
7. Discussion Extraction

   We use REST API to pull authoritative Tweets with the additional parameter of looking for tweets with ID bigger than the max ID. Afterwards we use the SEARCH API to find discussion. Then we examine all the Tweets collected in the 7 days for new discussion Tweets, using the MaxID and Search API. We observed that it is rare for a single handle to produce 3200 Tweets in one day. As a result, running the cycle once per day to collect the tweets posted can be enough. More frequent cycles can be run based on the problem tackled. There is a relatively low rate of Tweets produced by any authoritative handle; on average of 30 tweets per week. We used the anonymous calls (150 per/hour x200 Tweets per pull) with the Twitter REST API. The processing steps described above are achieved by the reference architecture illustrated in figure 1. The results from the search function referred to as ''discussion tweets" are compared to the original tweet used to create the query. The algorithm then compares the original tweet to the discussion tweet and identifies any information added to the original tweet. At this stage we focused only on the case where additional information is placed at the beginning or at the end of the tweet. This additional data will then be stored as the expression to be analyzed further to determine the sentiments of the additional information added by a public user. It should be pointed out that this module was developed for detecting the parts added to an authoritative Tweet, which is particular to the challenge we are addressing.

4. ADDITION DETECTION

The Difference algorithm compares two pieces of text and detects what should be added or removed to one to be identical to the other. Table 1 illustrates the output of the modified difference algorithm that we applied on Tweets. Two Tweets act as input; the first is the original Tweet T_1 the other is T_2. When an authoritative Tweet is detected, the Tweets are scanned for the same #tag and the original author. This signals that the Tweets are similar and can be candidates for further investigation for discussion Tweet detection. The two Tweets are supplied to the

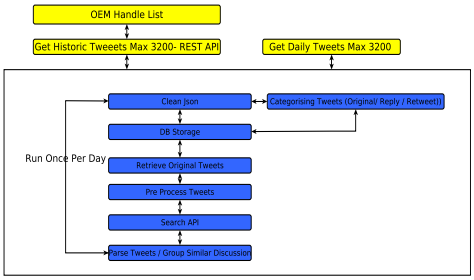


Figure 1: Reference Architecture and Overview.

Difference module that detects what needs to be added to the original Tweet to be the same as the discussion Tweet. We also calculate a similarity measure which is a normalized count of the word intersection between the two Tweets. The number is normalized to the total number of words in the shorter Tweet. Table 1 illustrates two Tweets; one of which is totally included in the other. Algorithm 1 lists the main steps for detecting additions to original tweets. The importance of addition detection can be summarized as an initial phase of collecting similar messages and further detect discussions and model emerging topics; [4] and [2] presented a method for topic model analysis of disaster-related twitter data. Methods presented depended on grouping tweets with similar hash tags. However, no semantic approach was developed to group semantically relevant tweets. Further, finding discussion tweets is an essential phase in detecting emerging topics and events.

Algorithm 1 Detecting Additions to the Original Tweet Psuedocode

1: **Input**:
- **Authoritative Tweet** T_1
- **Tweet with the same hash tag like the authoritative** T_2

2: **Output**:
- Added parts to the original Tweet

3: Initialize(T_1);
4: Initialize(T_2);
5: CalculateSimilarityPercentage(T_1,T_2);
6: CalculateNormalisedSimilarityScore;
7: CalculatePartsToBeAdded(Result);

5. DESIGN DECISIONS & CHALLENGES

This section presents the design decisions and the pitfalls that we encountred. We explain how we worked around the pitfalls. There were two options for querying the REST and the Search API: either JSON or XML. We decided to query the Twitter REST and Search API with request being made in JSON format, since there is already native support from PHP for parsing the JSON format. We chose the REST API since it enables us to access historical Tweets which is not possible with the Twitter Streaming API.

We used the ''GET statusesuser_timeline" HTTP request method from the REST API which we used to access a maximum of the last 3200 Tweets (including re-tweets) of an authoritative Twitter account. This function was chosen because it does not limit the owner of the Twitter account (i.e. the authenticated user).

Tweet Number	Tweet Content	Results when sent as a Query
1	Fires `http://t.co/KA8OOTNL`	No Results
2	Fires	Gets several results.

Table 2: Example of No Result for a Tweet with a Short URL

The Search API function ''GET search" was used to identify in public Twitter accounts the following messages:

1. Replies to the original Tweet (made by using the ''Reply" Twitter button)
2. Unofficial re-tweets (re-tweets not created via the Twitter ''ReTweet" button)
3. Tweets which mention the original Twitter's handle and some of the original text, but were not created with the ''Reply" or ''Re-Tweet" Twitter buttons.
4. Verbatim Tweets which are created using the ReTweet button - this is just a count of times the verbatim Tweet was re-shared, because you cannot both use the ReTweet button and modify the Tweet.
5. Verbatim copying with no reference to the source of the original Tweet (i.e. Messages are identical from different Twitter handles, but have no additional information to identify who is the original author).

Each of the Tweet types identified above (see list) can refer to the original Tweet in a variety of formats, each associated with different types of metadata in JSON query. The program is written in PHP with no web interface. All the data are stored in .txt files which form an exhaustive record. Only relevant data and metadata are then pushed into a MySQL database. PHP was chosen for its ease of use and installation, access without third-party platform, prior knowledge and experience in the team. MySQL is a free, open source supported database that interfaces well with PHP.

5.1 No Results, Search API

In the Search API documentation it states that the terms entered in the search box are searched by doing an 'or' among the words that match the query. However, the SEARCH API deals with shortened URLs inconsistently. For example, searching for ''fires" and a shortened URL provides no results, but removing the shortened URL and searching for ''fires" provides lots of results; this is illustrated in table 2. It is possible that shortened URLs are not indexed. As a result, we remove any URL from the query before supplying it to the Search API.

5.2 Tweet Truncation

Twitter stores long messages truncated within their database. The metadata can be used to check if the Tweet was truncated.

Tweet Author (OEM Handle)	**Original Tweet (T_1)**	**Discussion Tweet (T_2)**	Data Added Before the Original Tweet	Data Added After the Original Tweet	Similarity Measure
COEmergency	Fires #sanbrunofire http://t.co/29agn5mM	Horrible Disaster Fires #sanbrunofire http://t.co/29agn5mM three more victims until now #sbf RT @COEmergency	Horrible Disaster	three more victims until now #sbf RT @COEmergency	100 %

Table 1: Example of Detection of Additions to Tweets

Tweet	**Flipped Tweet Text**
San Bruno fire is a disaster #SBfire	#SBfire san bruno fire is a disaster

Table 3: Example of Flipped Tweet

Tweet	**Flipped Tweet Text**
1	LAFD: *Auto into Structure* 1828 E Marengo St x N Clement St; MAP 635-A3; FS 2; Auto into restaurant.... Read more at http://t.co/KA8OOTNL
2	RT LAFD: *Auto into Structure* 1828 E Marengo St x N Clement St; MAP 635-A3; FS 2; Auto into restaurant.... Read more at http://t.co/zjGOZYRa

Table 4: Example Tampering with Short URL's for Re-tweets

5.3 Backward Tweets

Table 3 shows an example of a backward Tweet. The inverted Tweets need to be detected since the encoding they use is different from the Latin encoding. In order to detect this format we excluded Tweets having a character set outside the Latin (a-Z).

5.4 Tampering With Short URLs for Re-tweets

The re-tweet feature can alter a shortened URL by Twitlonger or bit.ly. When re-tweeting a shortened URL, the shortened URL is given a new abbreviation. This only happens occasionally. As a result, link resolution can be a good solution to detect the original long term URL and match it; table 4 illustrates an exampled of tampered tweets.

5.5 Non Latin Characters

We are interested in collecting English based Tweets. To collect Tweets with a different character set the database should be changed in order to accept such a character set. Further, the meta data existing in the Tweet JSON should be checked for collecting Tweets authored in a specific languages.

5.6 Standard Language from Facebook

Some of the original Tweets have very popular text. Table 5 shows an example of a very popular Tweet that when used to query the SEARCH API gets a lot of results that were irrelevant. When we investigated this case further we found out that Facebook has standard text for posting a picture on Facebook. In order to avoid getting irrelevant results we filter Tweets that has the handle name and the hash tag of the authoritative Tweet.

Tweet	**Original Tweet Retrieves**	**Sample Results for Discussion**
ABC Emergency Mgt	I posted a new photo to Facebook http://t.co/6vJJplc7	I posted a new photo to Facebook http://t.co/1ZW64CP4
ABC Emergency Mgt	I posted a new photo to Facebook http://t.co/6vJJplc7	I posted a new photo to Facebook http://t.co/VBq6B3P0

Table 5: Standard Language from Facebook

5.7 MALFORMED JSON

Some of the fields in the JSON are not enclosed by quotations, which resulted that the native JSON decoder in PHP broke. As a result we developed a parser that adds the proper quotes whenever they are missing. We observed that this was more frequent with Boolean values. The customized parser took care of the malformed JSONs; as a result, no more exceptions are thrown by the parser.

6. DATA COLLECTED

From the Tweets collected, we record the following data which we use for further analysis: Number of times a verbatim Tweet is re-tweeted. This is a measure of popularity or positive support for this message. We also collect text on the right, in the middle, or left side of the original Tweet. Semantic analysis of this text can provide information on the user's opinion regarding the quality of the Tweet. Also the method of propagating the message (e.g., RT, @mention) is recorded. This indicates the strategy of message propagation. We are recording Tweets that contain verbatim or partial text from authoritative Tweets but do not include the authoritative handle of the original Tweet. These can be excluded from the analysis with post processing. These Tweets may inform the related discussion in the Twitterverse outside of authoritative Tweets.

7. CONCLUSION & FURTHER RESEARCH

We presented a framework to collect Tweets. The framework is general enough to be used the data collection needed from Twitter. We are sharing our pitfalls, experience and we will

share our code with the research community to benefit from our work. We pointed out the caveats that we faced and how we dealt with them. Current undergoing work involves clustering the messages collected according to their semantic relatedness, latent semantic inferences, non negative matrix factorization and Wikification relatedness. The current undergoing work will be used to enhance discussion detection and discovery. We are studying publishing the services of crawling and clustering to the community and making it accessible as a webservice in the cloud to reduce the time needed to collect Tweets and discover meaningful relations among tweets.

Future work will involve techniques to collect and discover discussions and emerging stories in Tweets. We are also incorporating an automated mail server that can email a set of accounts results that are of interest regularly. Further research will delve in automatic event detection.

We focused on scaling the techniques developed. Further experiments for discussion collection and emerging event prediction of the modules are crucial to explore their ability to provide prediction and clustering. The event detection can help in increasing safety and awareness towards an event in Another aspect for future research is dealing with updates. Incorporating different types of updates is an interesting challenge that needs to be taken into consideration to build an efficient way to collect discussions and new Tweets.

8. ACKNOWLEDGEMENTS

This work has been supported in part by the CMUSV Disaster Management Initiative and by a DMI affiliate, IntraPoint. The work has also been supported by IMT Lucca Italy. We thank Dr. Martin Griss and Dr. Patricia Collins for advice and multiple reviews of the paper, and their many comments and changes that added to the quality of this paper.

References

[1] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768. ACM, 2010.

[2] S. Doan, B. Vo, and N. Collier. An analysis of twitter messages in the 2011 tohoku earthquake. *Arxiv preprint arXiv:1109.1618*, 2011.

[3] A. Hughes and L. Palen. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6(3):248–260, 2009.

[4] K. Kireyev, L. Palen, and K. Anderson. Applications of topics models to analysis of disaster-related twitter data. In *NIPS Workshop on Applications for Topic Models: Text and Beyond*, 2009.

[5] S. Kumar, G. Barbier, M. Abbasi, and H. Liu. Tweettracker: An analysis tool for humanitarian and disaster relief. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[6] R. Munro. Subword and spatiotemporal models for identifying actionable information in haitian kreyol. In *Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011)*, 2011.

[7] E. Myers. An o (nd) difference algorithm and its variations. *Algorithmica*, 1(1):251–266, 1986.

[8] S. Vieweg, A. Hughes, K. Starbird, and L. Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 1079–1088. ACM, 2010.

Cluster Analysis on User Profile Variables for a Digital Media Learning Environment

Arturo Fernandez Espinosa, Meaghen Regts, Jayshiro Tashiro and Miguel Vargas Martin
Faculty of Business and I.T., University of Ontario Institute of Technology, Oshawa, Ontario, Canada

Abstract— *Neural Networks have been widely used to perform different tasks related to diverse scientific areas. Neural networks are a powerful tool that deal with the linear separability problem. In this paper the study objects are visualized as high dimensional vectors. These are separated according to their characteristics using diverse neural network architectures in order to compare results for a prediction model that has as intention to predict the final grade of the student within a Digital Media Learning Environment. We use population of 69 users, each of them represented by eight variables plus a target variable called grade. The actual work shows a comparison between neural network architectures to predict the target variable grade based in the eight variables that conform the high-dimensional vector.*

Keywords: Data Mining, predictive models, neural networks, knowledge discovery, feature extraction.

1. Introduction

Prediction models are an important study area in different fields of science, predictions are use everywhere in our daily life. From weather prediction, to the stock market, from prevention of natural disasters to gambling decisions in big casinos. Prediction models can be applied to any single aspect. Decision trees, linear regression, neural networks, support vector machines to mention some of the well known prediction models algorithms...

The present work will exploit those capabilities coupling a prediction model with a neural network.

The neural network will use as an input data the high dimensional vectors of 13 variables. We have 69 cases, 75% of then will be used as a training set and 25% as testing cases to check the accuracy of the model. The target or output variable will be the grade in the activity for each user which is a numeric variable that goes from 0 to 18 according to the student performance within the activity in IPSims.

Different models will be tested in order to determine which neural network architecture show the most accurate results for our study cases. The rest of the paper is organized as follows: Section II will show related literature about prediction models and neural networks. IPSims will be briefly described in Section III. Section IV will show the selection methodology of our prediction model. Section V presents the conclusions and final remarks of this work.

2. Related Work

Data mining applied to the education is an area that have been growing lately for the impact that it represents in the educational institutions. Multiple works have been presented in the educational data mining having as main objectives detect factors that can affect the student success within his education in different levels, generate better teaching techniques or tools based in the extraction of knowledge from large educational databases, etc.

In the large amount of works related with educational data mining we can find some of them with similarities to our work. Sembiring et. al. [1] Shows a similar process to the one performed in our research, cluster analysis in order to detect some correlations and specific groups of students based on a criteria, in the case of this paper the clusters are classified according to the performance of the students in a high learning institution. The final goal is to predict which students will be: Excellent, very good, good, average and finally poor. This will help to detect why a lot of students drop out, or why we count with genius according to a group of six variables applied to 1000 students in a form of questionarie. The authors apply a kernel k-means in order to generate the clusters of students after that they use smooth support vector machine to generate a prediction model. In our case k-means were apply to generate the clusters and eventually a predictive model based on neural networks was used. A big disadvantage in our work is the lack of study cases counting just with 69 in comparison the 1000 in the presented paper by Sembiring et. al.

Pardos et. al.[2] Is another description of a prediction model for the educational area, this work is peculiar because the authors focus more in the level of granularity that will make a bayesian prediction model more accurate. They use more than 600 students using levels of granularity that goes from 1, 5, 39 and 106 skills. The authors want to predict the final grade of the student within the general Massachusetts standardized state test. And they conclude that the prediction model based on bayesian networks is better when they have a bigger number of skills described in the bayesian network. We can see once again the weak point of our work with the

number of study cases, they count with 600 students while we just have 69.

Kabra et. al.[3] Shows another similar work where the prediction model is applied to engineering students but in this specific case using decision trees. The intention on this work is to predict who are likely to fail and with this information allow the teacher to provide feedback in order to avoid the undesirable result. This study count with 369 study cases all of them students of engineering. This study is a prove that knowing the past academic story of the students we are capable to generate an accurate prediction model to identify the students that are more likely to fail engineering courses.

Ramaswami et. al. [4] propose the CHIAD prediction model which is applied over 772 students of higher level education. A questionarie was applied to the students as well some other information were recollected from the students office, this information conformed a 35 variables dataset that was analyzed by the proposed algorithm CHIAD which is a classification tree algorithm. The algorithm prove to be an efficient tool to determine rules and predict the success of the students based in the 35 variables, as well a subset of stronger impact variables were identified. It is important to notice the similarity in the approach of detecting variables with higher impact in our past work for cluster analysis [6].

3. IPSims Description

Digital Learning Media Environments have been designed to support distinct activities related with learning and apprenticeship. IPSims was designed with the intention to support the learning activities of health sciences students as well as a provider of valuable information for researches in the study of misconceptions and student behaviours within computer based simulations.

IPSims present a complete source of information, videos, medical profiles, local and external literature resources presented in an efficient way to the student. IPSims is a complete platform for learning support and research purposes. The complete environment is generated by a template created with FLEX technology [5] .

The first encounter of a user with IPSims is as described in Figure 1 . This is the main form where the user has to fill the form in order to generate a user profile. After fill the required information and type a desired password the user will click the the register button. IPSims then will generate a user ID that will be the identification of the user inside the system.

IPSims is built-in with different tools that provide different functionalities. The activity of each user within the system is tracked by what we call path-finder. This tool allows the researchers to observe the decision sequela of each user as well as the time stamps. Each hyperlink triggers an event with a page ID and the time stamp, eventually this is store in a Postgress database. This Postgress database is the information core where the user profiles, decision sequela and time stamps, as well as information collected from a questionary are stored here to be extracted and analyzed later.

In this paper just the information from the user profile coming from the main screen as well some variables collected from a questionary will be the ones that will be used as an input for the prediction model. The criteria used for the selection of these variables is described in the next section

4. Data Collection

The strategy to collect the data was planned to be within a couple of courses in health sciences. The students from the class had an information session to show them how to use the system, as well a paper questionary was given to them. This questionary were related with a user profile and usability questions as well another questionary was handled in order to evaluate their performance within the activity.

After prepare the students to use the system. They filled the paper questionary with the questions related to their user profile. Then they proceed to generate their user profile Figure 1 to access the system in their computers. They were required to accomplish with the activities that conformed one of the scenarios presented on IPSims and they have to answer a paper questionary to evaluate their performance within the activity Figure 2.

69 students were recruited within the two different courses of health sciences. The information was treated and cleaned from inconsistencies and outliers. All the variables from the paper questionary were persisted into a database that eventually was merged with the existent information in the system related with the user profile generated in the main screen of IPSims.

5. Variables Selection

The input set for the prediction model is being selected from the main screen of IPSims variables plus some of the variables collected from a questionarie applied to the students. The original set of variables is big in comparison with the number of students. This original set was reduced based on the factor analysis dimensionality reduction technique , which select a subset that will keep the an acceptable level of features to be used as an input for the prediction model.

After applying factor analysis reduction we obtained a Kaiser-Meyer-Olkin measure with a value of 0.541 which means that the feature extraction results will be acceptable. According to the literature values from 0.5-0.7 are mediocre, 0.7-0.8 good, 0.8-09 excellent and greater than 0.9 will reflect superb results in the feature extraction. Our value is in the acceptable level but will retrieve mediocre results in the feature extraction. Once analyzed the correlation matrix, component matrix, as well as eigenvalues and variances the final subset of variables is selected.

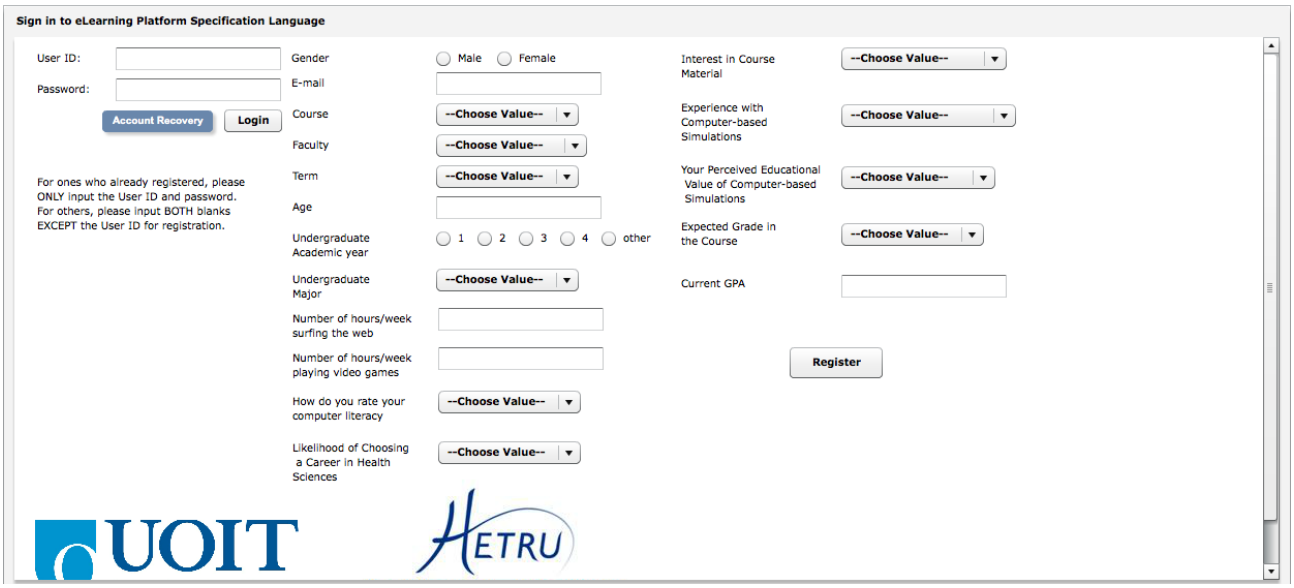


Fig. 1: IPSims Main Screen

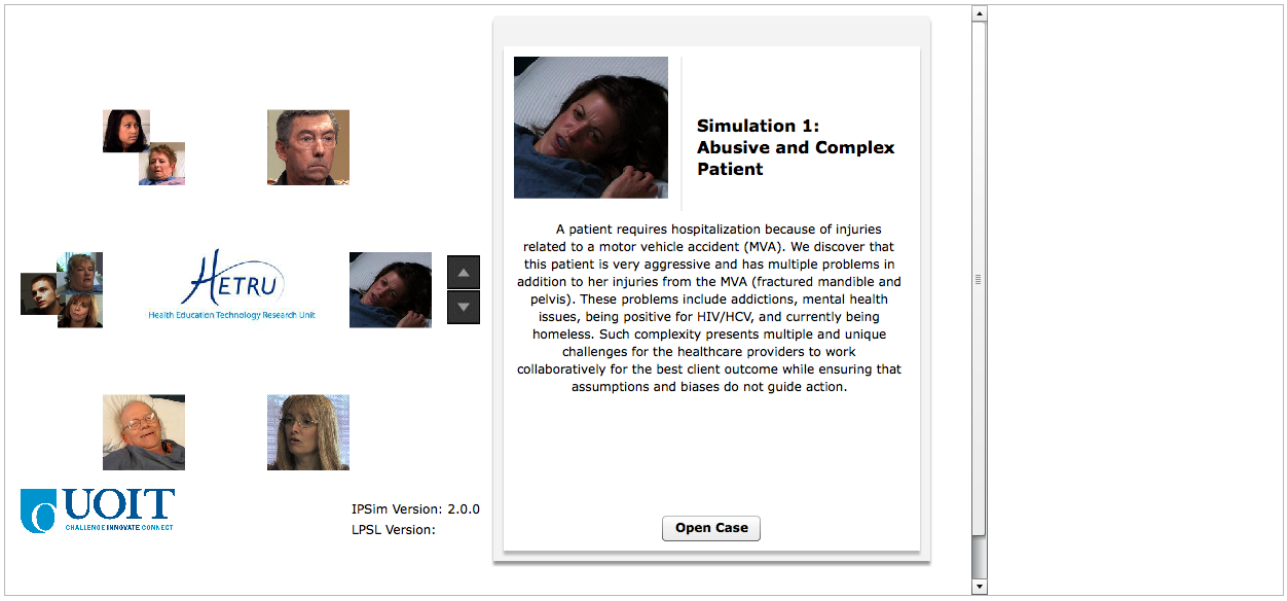


Fig. 2: IPSims main menu

The final subset of variables taken from the IPSims database and the questionary after the factor analysis for dimensionality reduction are as follows: Surf hour per week, high school average, pursue of post secondary education, age, gender, computer literacy and likelihood of choosing career in health sciences.

These selected variables are expected to be a good input for our predictive model. If we reach acceptable results based on measures as the R-squared and maximum absolute error as well as output plots we will be proving that this subset of variables is a subset that will reflect a direct impact in the performance of the student with a computer based simulation activity and we will be able to predict the final grade of the student based just in the user profile. This will give us the choice to give feedback or try to apply some techniques in order to improve the performance of a given group of students.

6. Prediction Models

Different architectures were tested in order to select the best network to predict the final grade within the activity based in the user profile. The proposed architecture for the

neural network was a graph with complete connectivity made of 8 input nodes with bias, a hidden layer with 6 neurons and one output. After generate this selected graph different training algorithms were tried.

The selected training algorithms were: conjugate gradient algorithm, global weights optimization with genetic algorithm and a distributed training algorithm. All the hidden layers use sigmoid function activation.

For the training set 51 one cases were selected for the training of the models and 18 for the testing, unfortunately some results were really poor and the predictive models give usefulness predictions but the distributed algorithm gave good results. The used tool to generate the models and training algorithms as well as the testing runs show different results depending on the run, different runs were made and different results were obtained in each run, some of them were really good, some others were acceptable and some others really poor, for each run a plot is generated in order to make visible the accuracy of our predictive model.

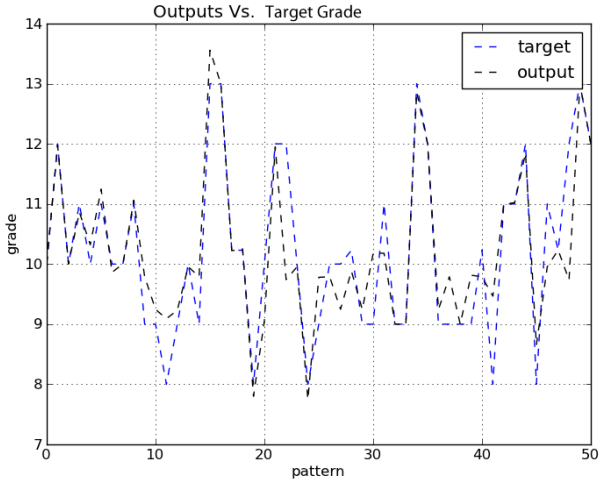


Fig. 3: Distributed algorithm training algorithm

7. Conclusions

The factor analysis result was acceptable giving us a Kaiser-Meyer-Olkin acceptable value, this had big impact in our experiments, trying to keep as much as possible the integrity of the information we generate a subset of variables that is more consistent and keep a low difference between the ratio of study cases vs. Variables. the ratio is from the order of 1:8.625 which is acceptable for the low number of study cases in our experiments. With just 69 study cases the training and testing of our predictive model could lead to trivial results.

Fortunately the results seem to show that with the dimensionality reduction analysis we obtain a valid dataset that combined with the study cases generates a well trained prediction model. These results suggest that the selected subset has a direct impact in the final grade of the activity for the user. The capability to predict the final grade based on the user profile is a powerful tool that open a window to new studies in the area education. The possibility to give some special attention or suggestions to the users that own a user profile with low expected grade as well identify the values which define the user groups according to their grades is valuable knowledge extracted from the original database.

The best results were shown for the distributed training algorithm, the R-squared values are in the average of 0.70-0.80 in the multiple runs where the algorithm were tried. The maximum squared error average is around 0.15-0.24. The results of one of the runs can be observed in Figure 3. The other two training algorithms shown poorest results. The conjugate gradient algorithm shows values around 0.23-0.35 for the R-squared measure Figure 4 shows a run of the algorithm. The genetic algorithm optimization training algorithm threw values of around 0.11-0.19 Figure 5 shows the results for one of the runs.

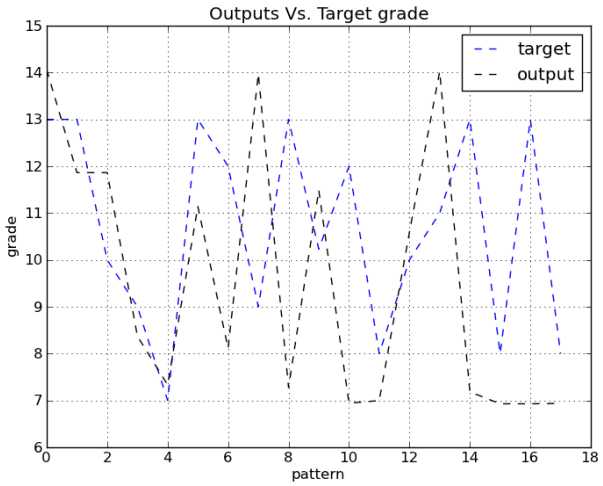


Fig. 4: Conjugate gradient training algorithm

Different architectures as well training set selection could be made and the best combinations could be selected by cross-validation this is the next stage on our work. The correct subset of variables have been found, the next stage is to generate an ideal model for prediction. The only concern now is that the study cases is really small, so there is not certainty that the model will work efficiently with different cases from the ones we have already for training and testing. Is important to recruit new study cases in order to increase our database and generate a better trained model as well have more testing cases in order to prove that effectively our model is working as desired.

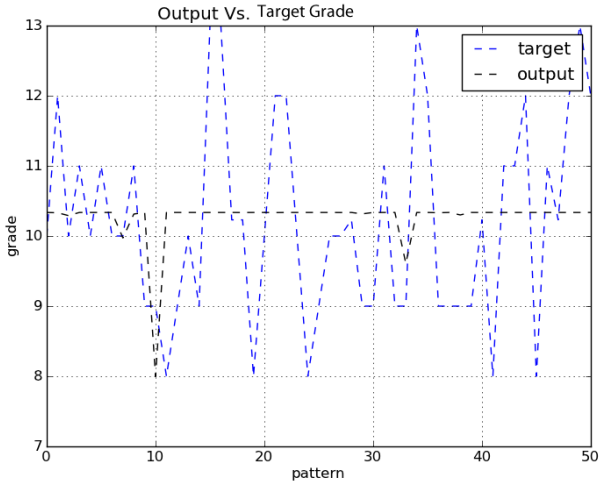


Fig. 5: Genetic algorithm optimization training algorithm

Acknowledgment

This work was supported in part by the Social Sciences and Humanities Research Council of Canada.

References

[1] S. Sembiring, M. Zarlis, D. Hartama, S. Ramliana, E. Wani, *Prediction of Student Academic Performance by an Application of Data Mining Techniques*, 2011 International Conference on Management and Artificial Intelligence, Vol 6, pp. 110-114.

[2] Z. Pardos, N. Heffernan, B. Anderson, *The Effect of Model Granularity on Student Performance Prediction using Bayesian Networks*, User Modeling 2007, Lectures Notes in Computer Science, Volume 4511, pp. 435-439.

[3] R. R. Kabra, R. S. Bichkar, *Performance Prediction of Engineering Students using Decision Trees*, 2011 International Journal of Computer Applications Volume 36 No. 11, pp. 9-12.

[4] http://www.adobe.com/devnet/

[5] M. Ramaswami, R. Bhaskaran *A CHAID Based Performance Prediction Model in Educational Data Mining*, IJCSI International Journal of Computer Science Issues, Vol 7, Issue 1, No. 1 January 2010 pp.10-18.

[6] A. Fernandez, M, Regts, J, Tashiro, M. Vargas Martin, *Cluster Analysis on User Profile Variables for a Digital Media Learning Environment*

Data Mining Utilization: A Successful Implementation Of Improved Clustering Algorithm Toward Identifying Crimes Patterns

Ahmed Alghamdi, Dr. Hanney Shaban
Department of Computer Science, College of Engineering, The Catholic University of America
10alghamdi@cardinalmail.cua.edu & shabanh@cua.edu

Abstract— **The purpose of this paper is to illustrate a data mining application method of clustering algorithm and its successful implementation toward identifying crimes patterns. Crimes are considered a community trouble and price our civilization extremely in several ways. To decrease the amount of crimes, data mining can be utilized using historically large data sets. Also, to identify the crimes samples and accelerate the procedure of resolving them, clustering algorithm for a data mining method is utilized. A k-means clustering with various improvements are taken into account, in order to help in the procedure of recognition of crime samples. Association rules are concerned on each year independently. Lastly, the rest of the essential statistic, pivot table and charts on each information group are joined. The combined information group is very practical, as well as it obviously specifies to the most unsafe regions in DC, time span and the frequent crimes in several regions.**

I. INTRODUCTION

In its simplest form, data mining refers to the process of discovering new patterns from existing large data sets in the form of meaningful knowledge. Accordingly, such reference implies the notion of hidden information in the available data yet to be discovered. Information technology as the enabler of such a process has rapidly grown since the innovation of computers. It dramatically revolutionized many aspects of organizations conducting their business among which is data warehousing; much data collection and storage and less utilization and analysis.

This paper attempts to illustrate four data mining jobs driven from the published crimes information set for 2006, 2007, 2008, 2009, and 2010 years by Washington DC government. In addition, these jobs are investigated that they are concerned on these information group independently and after these years are combined together.

Several jobs that are concerned on these information groups, which are Clustering, Classification, Correlation Analysis and Association Rules. Additionally, to recognize the crime information, various essential statistic, cross-tabulation, and charts have been concerned. Initially, these techniques are concerned on each information group and then all information has been combined from 2006 to 2010 and is stopped with about 170,000 records. This project is intended to discover the concealed sample and the associations between all aspects such as crime=burglary and season= summer, in order to create a form to expect a class aspect as well as put these enormous quantity of record within clusters.

a. In Classification, in 2006 a class aspect have been built information group that inquires whether the crime appears to be planned or not, which indicates that not only the usual arrangement come before approximately all crimes, but this indicates that the crime appears to be made by a specialized unlawful, or by a set of unsafe criminals, which has undiscovered features. Conversely, crimes that are made by person or do not appear to be made by set of criminal or unsafe criminals, are presented. The values have been penetrated of this aspect by hand after the information is investigated cautiously and set up out laws, in order to count on the crime, as well as on the crime technique explanation in the Meta data that has been published by the DC government. Subsequently, a form is built to be utilized for future illustration, when the criminal disappeared or the crime is reported adjacent to unidentified.

b. In Association Rules, the previous algorithm have been concerned on each year independently, then it is concerned to all years that are combined together to discover the concealed rule and compare them.

c. In Clustering, each year has been gathered into 10 clusters, and then all year information is gathered within 20 clusters. The clustering would permit us to expect the future year crime particularly if the crimes have replicated.

d. In correlation, 20 aspects have been gathered out of 94 aspects carefully, in order to discover the negative and positive associations. Even though, any exciting association is not discovered, but in fact, no important association is fine enough.

e. Lastly, the rest of the essential statistic, pivot table, and charts on each information group are joined. The combined information group is very useful and obviously designates to the most unsafe areas in DC, time span, and the frequent crimes in several areas.

The paper is organized as follows: the next section presents basic statistics about crimes from the collected data sets. The following section illustrates the used clustering technique. A correlation analysis is then provided. The final section provides a summary and notes some restrictions regarding the application of the presented data mining technique.

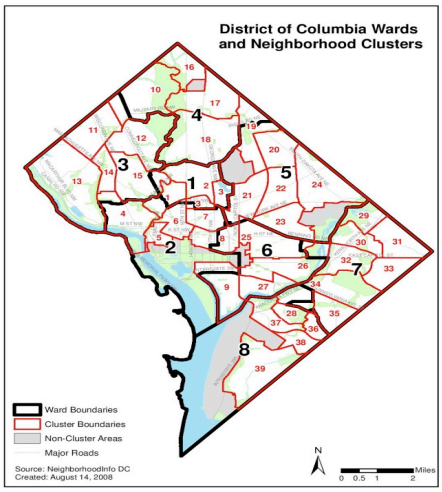


Fig. 1. District of Columbia Wards and Neighborhood Clusters

Figure 1 illustrates an example of map that it is useful to look and recognize this kind of map, before reading the investigation to be able to recognize great branch of information that communicates the places in the information such as WARD and NEIGHBOORHOOD CLUSTER aspects.

II. Basic Statistic of Crimes from 2006 to 2010

All data have been integrated into one file to do some statistic and data mining technique in order to see how all data together in comparison to single year data (170,000 crime records).

TABLE 1
NUMBER OF CRIMES BY SEASON

Season	Winter	Spring	Summer	Fall
Number of crimes	37028	44398	45633	41894

TABLE 2
NUMBER OF CRIMES BY TIME

Time	PM	AM
Number of Crimes	96558	72395

TABLE 3
TOP FOUR OFFENSES

Offenses	THEFT	THEFT F/ AUTO	STOLEN AUTO	ROBBERY
Number of Crimes	44291	39954	26260	21621

TABLE 4
TOP THREE METHODS

Methods	Not Occupied Building, Vehicle, Store	Occupied Building, Vehicle, Store	STOLEN AUTO
Number of Crimes	66146	29616	26260

TABLE 5
TOP THREE DANGEROUS WARD

Ward	WARD2	WARD1	WARD6
Number of crimes	31370	24559	23632

TABLE 6
TOP THREE DANGEROUS NEIGHBORHOOD CLUSTERS

Neighborhood	Neighborhood Cluster 2	Neighborhood Cluster 6	Neighborhood Cluster 8
Number of crimes	13910	10895	10472

III. Clustering Techniques Used

Twenty clusters have been created to get focused sight of 170,000 crime records over the period from 2006-2010. These clusters will be summarized based on the Neighborhood clusters number

a. Neighborhood (#1) has appeared in only cluster 18 that represents only 1% (2398 crimes). Neighborhood 1 belongs to ward 1, the season is summer and the shift is PM. The offense is THEFT AUTO; the method that is used to do this crime is "occupied vehicle".

b. Neighborhood (#2) has appeared in 4 clusters, three seasons appeared in these clusters. Winter season has two different offenses in this cluster, which are Burglary in AM shift, and THEFT AUTO in PM shift. Both of these offenses happened in no occupied place or vehicle. In summer season, the crime type is robbery in the PM time, and the method is F&V. Finally in FALL season, not occupied vehicle is THEFT in PM time.

c. Neighborhood (#3) has been shown in cluster 12, cluster 12 seasons is fall, the shift is AM and the

crime type is THEFT of not occupied building or store. Neighborhood3 is part of ward 1.

d. Neighborhood (#6) appeared in clusters 2, the crime happened in spring season in AM shift. The type of crimes is THEFT of not occupied building, store or vehicle.

e. Neighborhood (#7) in cluster 14 only, the crime happened in the AM shift of FALL; the type of crimes is robbery and the method F&V.

f. Neighborhood (#8) in cluster 11 only. The time is AM of the winter; the crime is THEFT of not occupied building, store or vehicle.

g. Neighborhood (#13) in cluster 15, the season of crime is spring in the PM shift, the offense is THEFT F/AUTO and the method is occupied vehicle.

h. Neighborhood (#22) in cluster 16, the season is spring in PM time and the offense is THEFT of not occupied building, store or vehicle.

i. Neighborhood (#25) in cluster 19, the crime type is THEFT F/AUTO of not occupied building, store or vehicle during PM shift of summer.

j. Neighborhood (#26) in cluster 1, the crime happened during the AM shift of spring, the type of crime is robbery using GUN.

k. Neighborhood (#34) in cluster 8 and 0. In cluster 8 the crime happened during the PM shift of FALL, the type of crime is Burglary of not occupied building, store or vehicle. In cluster 0, the crime happened in fall during the AM shift, the type is THEFT and the method is not occupied vehicle.

l. Neighborhood (#38) appeared in cluster number 17, the crime in this Neighborhood happened in the AM shift of the spring season. The type and the method of offense is STOLEN AUTO.

m. Neighborhood (#39) has appeared in three clusters. The first cluster is 4, the crime of this cluster ADW, the season of this crime in Neighborhood 39 is spring during the AM shift, and the method of this crime has been entered as OTHER, which means none of the previously known method has been used in this cluster. The second cluster is 9, where the crime and method is STOLEN AUTO; the season is summer during the PM shifts. In cluster 13, the crime is THEFT of not occupied building during the Fall PM shift.

TABLE 7
SAMPLES CLUSTERS

20 Clusters	Season	Time	Offense	Method	Ward	District	Neighborhood Cluster
Cluster 0 13538 (8%)	FALL	AM	THEFT F/AUTO	Not occupied building, vehicle ,store	7	SIXTH	34
Cluster1 10950 (6%)	SPRING	AM	ROBBERY	GUN	6	FIRST	26
Cluster2 22232 (13%)	SPRING	AM	THEFT	Not occupied building, vehicle ,store	2	SECOND	6
Cluster3 10374 (6%)	WINTER	AM	BURGLARY	Not occupied building, vehicle ,store	1	THIRD	2
Cluster4 7301 (4%)	SPRING	AM	ADW	OTHER	8	SEVENTH	39
Cluster5 8341 (5%)	SUMMER	PM	ROBBERY	F&V	1	THIRD	2
Cluster6 9920 (6%)	FALL	PM	THEFT F/AUTO	Not occupied building, vehicle ,store	1	THIRD	2
Cluster7 3981 (2%)	WINTER	PM	THEFT F/AUTO	Not occupied building, vehicle ,store	1	THIRD	2
Cluster8 8879 (5%)	FALL	PM	BURGLARY	Not occupied building, vehicle ,store	7	SIXTH	34
Cluster9 10146 (6%)	SUMMER	AM	STLEN AUTO	STOLEN AUTO	8	SEVENTH	39

IV. ASSOCIATION RULES USED

The most interesting rule is CRIME-SEASON=FALL 41894 ==> TIME=PM 30572 conf:(0.73).

Crime Season FALL implies PM Time. In other words, 30572 out of 41894 crimes in FALL 2006, 2007, 2008, 2009, and 2010 happened in PM shift.

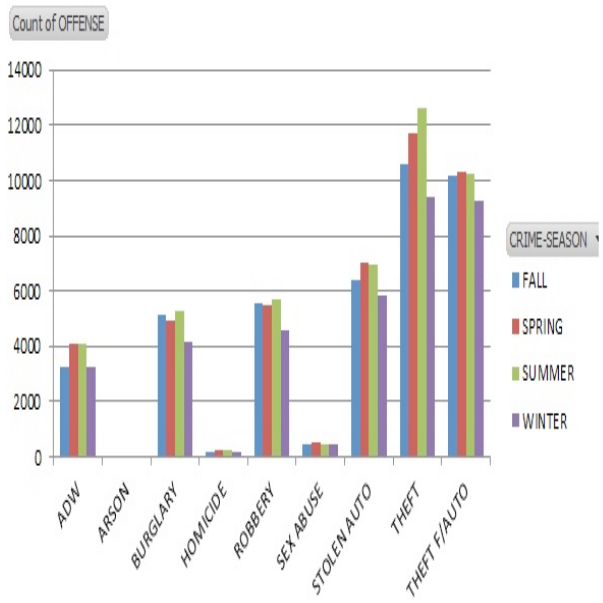


Fig. 2. the relationship between offense and season attributes

V. Correlation Analysis

The following correlation matrix is 20X20, some attribute have been omitted as they are not so important to this correlation analysis and because if all the attributes are included, the final correlation matrix is a matrix of 94X94 which equals to more than 4000 correlation. For example, the Neighborhood Clusters (values from 1-39) have been excluded, since ward (values from 1-8) is used to determine the place where the crime has been happened. Each ward actually contains several clusters, so this roiling up process will not affect this correlation analysis. Furthermore, Methods attributes (values 1-41) are excluded because this attribute is only further details to the crime type.

TABLE 8
CORRELATION ANALYSIS RESULTS

	WINTER	SPRING	SUMMER	FALL	HOMICIDE	ADW	BURGLARY	THEFT F/AUTO	STOLEN AUTO	ROBBERY	THEFT	SEX ABUSE	ARSON	WARD=1	WARD=2	WARD=3	WARD=4	WARD=5	WARD=6	WARD=7	WARD=8
WINTER	1																				
SPRING	-0.3163	1																			
SUMMER	-0.3223	-0.3632	1																		
FALL	-0.3042	-0.3428	-0.3493	1																	
HOMICIDE	-0.0028	0.0014	0.00164	-5E-04	1																
ADW	0.00058	0.01	0.0068	-0.018	-0.02143	1															
BURGLARY	-0.0057	-0.0076	-0.0008	0.0141	-0.02507	-0.11	1														
THEFT F/AUTO	0.01676	-0.0054	-0.0181	0.008	-0.0386	-0.17	-0.201	1													
STOLEN AUTO	0.00256	0.0053	-0.0041	-0.004	-0.02976	-0.13	-0.155	-0.24	1												
ROBBERY	-0.0049	-0.0043	-0.0024	0.0116	-0.02632	-0.12	-0.137	-0.21	-0.1628	1											
THEFT	-0.0107	0.0018	0.01989	-0.012	-0.04135	-0.18	-0.215	-0.33	-0.2557	-0.22615	1										
SEX ABUSE	0.00184	0.0061	-0.0045	-0.003	-0.00743	-0.03	-0.039	-0.06	-0.046	-0.04066	-0.06	1									
ARSON	-0.0005	-0.0033	-0.002	0.0059	-0.0025	-0.01	-0.013	-0.02	-0.0155	-0.0137	-0.02	-0	1								
WARD=1	0.01094	0.0066	-0.0069	-0.01	-0.00915	-0.02	-0.026	0.084	-0.0399	0.035719	-0.04	-0.01	-0.007	1							
WARD=2	0.00717	-0.0061	-0.0012	0.0006	-0.02651	-0.07	-0.047	0.033	-0.1175	-0.05698	0.201	-0.03	-0.013	-0.2	1						
WARD=3	0.00097	-0.0043	-0.0033	0.0069	-0.01463	-0.06	0.0075	0.031	-0.0627	-0.05726	0.103	-0.01	-0.008	-0.1	-0.11	1					
WARD=4	0.00424	-0.0046	-0.0021	0.0028	-0.00425	0.01	0.0035	-0.02	0.0234	0.025089	-0.03	0.006	0.0006	-0.13	-0.15	-0.08	1				
WARD=5	0.00316	0.0014	-0.0053	0.001	0.0064	0.02	0.0091	-0.01	0.0438	0.003992	-0.05	0.01	0.0047	-0.16	-0.19	-0.09	-0.12	1			
WARD=6	-0.0018	-0.0031	0.00938	-0.005	-0.0082	-0.02	-0.003	0.028	-0.0239	-0.00174	0.011	-0.01	0.0001	-0.17	-0.19	-0.1	-0.13	-0.16	1		
WARD=7	-0.0158	0.0077	0.00642	0.0008	0.017875	0.05	-0.002	-0.05	0.1196	0.012731	-0.1	0.015	0.0032	-0.16	-0.18	-0.09	-0.12	-0.15	-0.15	1	
WARD=8	-0.0099	0.0014	0.00204	0.006	0.038487	0.09	0.07	-0.1	0.0631	0.031007	-0.1	0.031	0.0198	-0.16	-0.18	-0.09	-0.12	-0.15	-0.15	-0.14	1

The correlation analysis ends up with no interesting results. All the results are very close to ZERO. The highest correlation that was found in the matrix is -0.36 between summer and spring attributes that means there is slight negative correlation between spring and summer crimes. Finally, the highest correlations that are found the matrix are highlighted in this matrix.

VI. Conclusion

The data mining for recognizing crime sample is utilized in this paper, which it use the clustering methods and Association rules. The decided donation was to create crime sample recognition as machine learning job and therefore to utilize the information mining to sustain police in decreasing crimes number. The important aspects are recognized. The specified modeling method is able to recognize the crime samples from a great amount of crimes that doing the police recognize where is the ward and neighborhood that require rising police patrol or establishing safety cameras to decrease amount of crimes there.

Several restrictions involve that crime sample investigation could be used only to assist the police, not to substitute them. The data mining is susceptible to superiority of input

information that might be imprecise; have absent data or being data access mistake level. Also, mapping actual information to data mining aspects is not constantly a simple job and frequently needs skilled information miner and crime information investigation with fine domain knowledge. They require operating personally with a police in the primary stages. As a future addition of this study, forms would be built for expecting the crime hot-spots that would assist in the operation of police at most probable places of crime for some known window of time, in order to permit the majority efficient use of police sources.

REFERENCES

[1] Washington DC Metropolitan Police Department at: http://crimemap.dc.gov/presentation/query.asp

[2] Introduction to Data Mining by Pang-Ning Tan, Yi Qin, Michael Steinbach, Vipin Kumar.

[3] Shyam Varan Nath. 2006. Crime Pattern Detection Using Data Mining. In *Proceedings of the 2006 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology* (WI-IATW '06). IEEE Computer Society, Washington, DC, USA.

[4] Bongjune Kwon and Hyuk Cho. 2010. Scalable co-clustering algorithms. In *Proceedings of the 10th international conference on Algorithms and Architectures for Parallel Processing - Volume Part I* (ICA3PP'10), Ching-Hsien Hsu, Laurence T. Yang, Jong Hyuk Park, and Sang-Soo Yeo (Eds.), Vol. Part I. Springer-Verlag, Berlin.

[5] M. A. Dalal and N. D. Harale. 2011. A survey on clustering in data mining. In *Proceedings of the International Conference & Workshop on Emerging Trends in Technology* (ICWET '11). ACM, New York, NY, USA.

[6] Nadim Asif, Faisal Shahzad, Najia Saher, and Waseem Nazar. 2009. Clustering the source code. *W. Trans. on Comp.* 8, 12 (December 2009), 1835-1844.

[7] Pinaki Mitra and Chitrita Chaudhuri. 2006. Efficient algorithm for the extraction of association rules in data mining. In *Proceedings of the 2006 international conference on Computational Science and Its Applications - Volume Part II* (ICCSA'06), Marina L. Gavrilova, Osvaldo Gervasi, Vipin Kumar, C. Kenneth Tan, and David Taniar (Eds.), Vol. Part II. Springer-Verlag, Berlin.

[8] Ruowu Zhong and Huiping Wang. 2011. Research of Commonly Used Association Rules Mining Algorithm in Data Mining. In *Proceedings of the 2011 International Conference on Internet Computing and Information Services* (ICICIS '11). IEEE Computer Society, Washington, DC, USA.

[9] Jochen Hipp, Ulrich G\&\#252;ntzer, and Gholamreza Nakhaeizadeh. 2002. Data Mining of Association Rules and the Process of Knowledge Discovery in Databases. In *Industrial Conference on Data Mining: Advances in Data Mining, Applications in E-Commerce, Medicine, and Knowledge Management*, Petra Perner (Ed.). Springer-Verlag, London, UK.

[10] Data Mining: Building Competitive Advantage by Robert Groth, *Santa Clara.*

A fuzzy clustering method using Genetic Algorithm and Fuzzy Subtractive Clustering

Thanh Le[1], Tom Altman[1], Katheleen J. Gardiner[2]
[1]Department of CSE, University of Colorado Denver, Denver, CO, USA
[2]Department of Pediatrics, University of Colorado Denver, Aurora, CO, USA

Abstract– *Clustering is a challenging problem in data mining, requiring both accurate determination of the number of clusters and correct clustering of the data. Fuzzy C-means (FCM) is a popular algorithm using the partitioning approach to solve this problem. A drawback to FCM is that it requires the number of clusters to be set a priori. In this study, we combine FCM with Genetic Algorithm (GA), Subtractive Clustering (SC) and Bayesian cluster validation for a novel clustering method, fzGASCE that both determines the correct number of clusters and efficiently constructs these clusters from a given dataset. We show that fzGASCE outperforms existing methods using similar approaches on both artificial and real datasets.*

Availability: *The test datasets and the method software are available online at http://ouray.ucdenver.edu/~tnle/fzgasce.*

Keywords: *fuzzy c-means, genetic algorithm, subtractive clustering, Bayesian cluster validity.*

1 Introduction

Clustering in data mining refers to the grouping of data points within a dataset based on their similar properties. Data points within a cluster are highly similar to each other and can be discriminated from data points within other clusters. Successful clustering, therefore, maximizes both the compactness of data points within a cluster and the discrimination between clusters. A clustering problem generally contains two parts: (i) determination of the number of clusters and (ii) assignment of data points to clusters. In the simplest application, the data are to be partitioned into only two clusters such that the summation of some non-geometric distance measures between the data points of the clusters is maximized. This clustering problem can be shown to be an NP-hard decision problem, for which there are currently no efficient algorithms that would run in polynomial time with respect to the dataset size [16]. Such problems are therefore "computationally intractable", thus allowing application of heuristic or approximation algorithms that provide results that may be adequate although not always optimal.

Clustering approaches include partitioning and hierarchical methods. Partitioning approaches include two methods for managing cluster boundaries: the well-known K-means algorithm that uses crisp boundaries and the Fuzzy C-means (FCM) [1] that uses fuzzy cluster boundaries, where fuzzy sets are applied to associate every data point with at least one cluster. The K-means algorithm requires that every data point belongs to only one cluster. It therefore may not be appropriate for some real applications, such as gene expression analysis or pattern recognition. On the other hand, the FCM algorithm, using the concept of fuzzy set theory to allow every data point to belong to more than one cluster, is more appropriate to real-world problems and it has become the most popular partitioning method. However, the FCM algorithm, like most other partitioning approaches, cannot by itself determine the number of clusters, and results depend strongly on initial parameters. For some initial values, FCM will converge rapidly to a global optimum, but, for others, it may become stuck in a local optimum.

One approach to address the limitations of FCM is to integrate FCM with the Genetic Algorithm (GA), where the GA is used to manage a set of solution candidates, the FCM algorithm is applied and a cluster validity index is used as the GA fitness function to search for the best solution. Ghosh et al. [7] and Liu et al. [10] proposed to use the partition coefficient (PC) [1] and Xie and Beni (XB) [14] validity indices. In addition, Liu et al. [10] proposed a modified version of the PC index (MPC) in order to reduce the monotonic tendency of the index. The Fukuyama-Sugeno cluster index (FS) [5], which measures the

This work was supported in part by the Vietnamese Ministry of Education and Training (TL).

Thanh Le is a doctoral student in the Department of Computer Science and Engineering, University of Colorado Denver, Denver, CO 80217, USA (email: lntmail@yahoo.com).

Tom Altman is a professor in the Department of Computer Science and Engineering, University of Colorado Denver, Denver, CO 80217, USA (email: tom.altman@ucdenver.edu).

Katheleen J. Gardiner is a professor in the Department of Pediatrics and the Linda Crnic Institute for Down Syndrome; and a member of the Intellectual and Developmental Disabilities Research Center; and the Computational Biosciences, Human Medical Genetics and Neuroscience Programs, University of Colorado Denver, Aurora, CO 80045, USA (phone: 303-724-0572; email: katheleen.gardiner@ucdenver.edu).

compactness and separation of the cluster partition, was used by Ghosh et al. [7]. Lianjiang et al. [8], in a novel self-Adaptive Genetic Fuzzy C-Means algorithm (AGFCM), proposed a validity index combining the PC index with total variation of fuzzy partition. Halder, Pramanik and Kar [6] proposed a GA fitness function (HPK) based on the compactness measure and combined it with an intra-inter validity index for a novel algorithm that automatically determines the number of clusters in the dataset. Lin et al. [9] proposed a combination of GA and FCM for a novel method with adaptive cluster validity index (ACVI) based on the intra and inter measures of the fuzzy partition, where GA is used to solve the trade-off problem between these two factors for a better evaluation of the fuzzy partition.

A common limitation of existing methods using GA with FCM is that the GA fitness functions are based on the cluster validity indices, which usually have a problem with scaling between the compactness and separation factors. In addition, they use the maximum membership degree for defuzzification that may be improper, because the membership is computed based on the distance between the data object and cluster center. Use of membership degree can assign marginal objects of a large cluster to the immediately adjacent small cluster. This is illustrated in Figure 1, where if a data object is in the gray rectangle, it may be incorrectly assigned to cluster 3 instead of cluster 2.

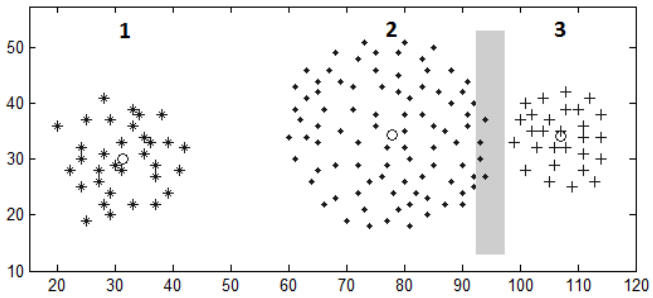


Figure 1: ASET4 - an artificial dataset with three clusters of different sizes

In this study, we combine FCM with GA and fuzzy SC algorithms for a novel clustering algorithm, fzGASCE, that automatically determines the number of clusters in the dataset. The FCM algorithm rapidly determines the exact clustering prototypes of each solution candidate so that the GA algorithm, managing a set of such candidates, can select the optimal one. The fuzzy SC method helps the GA algorithm escape any local optima.

2 Methods

2.1 Fuzzy C-Means algorithm (FCM)

Let $X = \{x_1, x_2,\ldots,x_n\} \in R^p$ be a set of data objects x_i, i=1..n. For a given c, $2 \le c < n$, the Fuzzy C-Means algorithm (FCM) divides X into c clusters by minimizing the objective function:

$$J_m(U,V) = \sum_{i=1}^{n}\sum_{k=1}^{c} u_{ki}^m d^2(x_i, v_k) \to \min \qquad (1)$$

where $u_{ki} \in [0,1]\ \forall k,i$,

$$\sum_{k=1}^{c} u_{ki} = 1 \forall i\ , \qquad (2)$$

and m, $1 \le m$, is the fuzzifier factor; V, $V = \{v_1, v_2,\ldots,v_c\}$ is a set of c cluster centers; $U = \{u_{ki}\}_{i=1..n,k=1..c}$ is a partition matrix; and $d^2(.)$ denotes the Euclidean norm.

Minimizing J_m with respect to (2), we obtain an estimated model of U and V as:

$$u_{ki} = \left(\frac{1}{d^2(x_i, v_k)}\right)^{\frac{1}{m-1}} \Bigg/ \sum_{l=1}^{c}\left(\frac{1}{d^2(x_i, v_l)}\right)^{\frac{1}{m-1}}, \qquad (3)$$

$$v_k = \sum_{i=1}^{n} u_{ki}^m x_i \Bigg/ \sum_{i=1}^{n} u_{ki}^m . \qquad (4)$$

FCM can converge rapidly and provide soft partitions applicable to many real-world applications. However, FCM depends strongly on initial parameters and can become stuck in a local optimum.

2.2 Genetic Algorithms (GA)

Genetic algorithms (GAs) are a class of optimization algorithms that perform adaptive searches to find solutions to large scale optimization problems with multiple local optima. When applied to cluster analysis, conventional GAs use chromosomes to describe solution candidates. Each chromosome represents a part or the whole of a solution. In the case of the later, a chromosome is a set of loci each standing for a data point selected as the center of a cluster in the solution. GAs manage a set of chromosomes to search for the best solution through an evolution process. At each generation, new offspring are created from the parents, chosen based on either tournament or roulette wheel selection methods, using the crossover and mutation operators. The worst members in the current generation will be replaced by the newly created ones to construct the next generation.

The crossover and mutation operators help GAs to escape local optima. These operators depend strongly on how the probabilities of crossover and mutation are chosen. Recent improvements in GAs have focused on adaptively adjusting operator probabilities so that the genetic processes rapidly escape local optima. However, setting up adaptive GAs is difficult and most approaches are based on heuristics.

3 The fzGASCE algorithm

We propose a novel fuzzy clustering algorithm, fzGASCE, which combines the GA and fuzzy SC algorithms with a Bayesian based cluster evaluation method to overcome the drawbacks of existing methods that use the GA and FCM algorithms.

Chromosome

We used chromosome to represent the whole clustering solution. Each chromosome contains a set of loci each standing for the index of the data point selected as cluster center. We set the length of chromosomes to $\sqrt{n}$, which is assumed the maximum number of clusters in the dataset.

Crossover operator

The crossover operator is used to produce two new offspring from a given pair of parents. Both the roulette wheel and tournament selection methods are used interchangeably to select parents maintaining potentially useful solutions in the current generation. A two-point crossover operator with probability P_c, $P_c = 0.5$, is used.

Mutation operator

The mutation operator is used to make changes in portions of the chromosomes of newly created members. Because each chromosome encodes data points representing the cluster centers of a clustering solution, changing of the data points in the chromosome may improve the clustering quality. We therefore propose three different tasks for the mutation operator: (i) add a data point as a new cluster center, because this may help to locate a new cluster in a higher density region, (ii) remove a data point to prevent the inclusion of a sparse cluster, and (iii) replace one data point with another so that the new cluster is located in a higher density region. These tasks are commonly used in existing methods. However, they are employed in a random or heuristic way, and cannot guarantee that the GA algorithm will escape local optima. To address this issue, we propose an alternative approach using the fuzzy SC method of Le et al. [13]. The fuzzy partitions of the parents are used to estimate the density at every data point.

$$\mathrm{dens}(x_i) = \sum_{k=1}^{c} \mathrm{Acc}(v_k) \times u'_{ki}\,, \tag{5}$$

where, $\mathrm{Acc}(v_k)$, the accumulated density at v_k, and $\{u'_{ki}\}$, a strong uniform fuzzy partition of $\{u_{ki}\}$ [2], are defined, respectively, for k=1..c, i=1..n as:

$$\mathrm{Acc}(v_k) = \sum_{i=1}^{n} u_{ki}\,, \tag{6}$$

$$u'_{ki} = \left[e^{d(x_i,v_k)/\sigma_k}\right]^{-1} \Big/ \sum_{l=1}^{c} \left[e^{d(x_i,v_l)/\sigma_l}\right]^{-1}. \tag{7}$$

At each time, t, the data point, x^*_t with the highest density, M^*_t, is selected, and the densities at the remaining data points are updated as:

$$\mathrm{dens}^{t+1}(x_i) = \mathrm{dens}^{t}(x_i) - M^*_t \sum_{k=1}^{c} u'_{ki} \times P(v_k \mid x^*_t). \tag{8}$$

The points where the densities in this way change less than a predefined ratio, R_M, are considered to be significantly dense and are used in the mutation operator of fzGASCE. This is instead of using randomly selected data points, as in existing methods. We chose a value of 0.95 for R_M. The selection of the value of R_M does not affect the outcome however a low value of R_M may slow the convergence process of fzGASCE.

Fitness function

Instead of using a cluster validity index for the fitness function, we use the method of Le et al. [12] for cluster evaluation. For each chromosome, FCM is applied to generate the fuzzy partition which is then used to generate a probabilistic model of the data distributions [3, 12]. For each cluster v_k, k=1..c, the probability distribution $\{p_{ki}\}_{i=1..n}$ is derived from the possibility distribution $\{u_{ki}\}_{i=1..n}$. Then, the following statistics at v_k are computed:

$$\sigma_k = \sum_{i=1}^{n} p_{ki} \left\| x_i - v_k \right\|^2, \tag{9}$$

$$P(v_k) = \sum_{i=1}^{n} P(x_i \mid v_k) \Big/ \sum_{l=1}^{c} \sum_{i=1}^{n} P(x_i \mid v_l)\,, \tag{10}$$

$$P(x_i \mid v_k) = \left((2\pi)^{1/n} \times \sigma_k \times e^{\frac{\|x_i - v_k\|^2}{2\sigma_k^2}} \right)^{-1}, \tag{11}$$

where σ_k and $P(v_k)$ are the variance and the prior probability of v_k respectively; $P(x_i|v_k)$ indicates the conditional probability of x_i given v_k, for i=1..n, k=1..c.

Given a fuzzy partition θ, θ = {U,V}, the fitness function is defined:

$$\mathrm{fitness}(\theta) = \log[L(\theta|X)] - \log(c), \tag{12}$$

where L(θ|X), the likelihood of the clustering model and the data, is measured as:

$$L(\theta \mid X) = L(U, V \mid X) = \prod_{i=1}^{n} P(x_i \mid U, V) = \prod_{i=1}^{n} \sum_{k=1}^{c} P(v_k) \times P(x_i \mid v_k). \quad (13)$$

Defuzzification of fuzzy partition

Use of the maximum fuzzy membership degree to determine classification of data points result in assignment of marginal objects of a large cluster to the immediately adjacent small cluster. The gray rectangle in Figure 1 shows an example where data points may be incorrectly assigned to cluster 3 instead of to cluster 2. Therefore, we use the probabilistic model. A data point x_i, i=1..n, will be assigned to cluster v_k, k=1..c, where:

$$P(v_k \mid x_i) = \max_{l=1..c} \{P(v_l \mid x_i)\}. \quad (14)$$

Because $P(v_k|x_i) = P(x_i,v_k)/P(x_i) = P(x_i|v_k)*P(v_k)/P(x_i)$, we solve this as:

$$P(v_k \mid x_i) = \max_{l=1..c} \{P(x_i \mid v_l) \times P(v_l)\} \quad (15)$$

***fzGASCE** algorithm*

- Input: data to cluster $X=\{x_i\}$, i=1..n.
- Output: an optimal fuzzy clustering solution,
  - c: optimal number of clusters.
  - $V = \{v_i\}$, i =1..c: the cluster centers.
  - $U=\{u_{ki}\}$, i=1..n, k=1..c: the partition matrix.

Steps

1. Randomly generate a set of chromosomes describing clustering solution candidates.
2. Compute the fitness value for every chromosome.
3. If the stop criteria are met, then go to step 7.
4. Apply the crossover operator with the probability P_c = 0.5, and the roulette wheel and tournament parent selection methods.
5. Apply the mutation operator with the probability P_m = 0.01. The significantly dense data points are used in the replacement, and fresh offspring are created using these points with a probability of $P_c \times P_m$.
6. Go to step 2.
7. Select the 'best' chromosome from the population as the clustering solution.
8. Apply (15) for the defuzzification of the fuzzy partition of the solution.

4 Experimental results

Datasets

To evaluate the performance of fzGASCE, we used four artificial datasets generated using an infinite mixture model method [15]. ASET1, ASET2 and ASET3 each have well-separated clusters of similar sizes. The number of clusters and data dimensions of these dataset are (5,2), (5,3) and (11,5), respectively. ASET4 is more complex, containing three clusters that differ in size and density (Figure 1). For real datasets, we used the Iris and Wine datasets from the University of California Irvine (UCI) Machine Learning Repository [4]. The classification structures in these datasets are known.

Performance measures

We used three measures, COR, EVAR and EMIS, to evaluate algorithm performance. COR is the correctness ratio, defined as,

$$COR = \frac{1}{N} \sum_{i=1}^{N} I(c - \hat{c}), \quad (16)$$

where N is the number of trials, c and $\hat{c}$ are the number of clusters and the predicted number of clusters respectively, and I(.) is defined as:

$$I(x) = \begin{cases} 1, & x = 0 \\ 0, & x \neq 0 \end{cases}. \quad (17)$$

EVAR is a measure of the accuracy of the predicted number of clusters defined as in (18).

$$EVAR = \frac{1}{N} \sqrt{(c - \hat{c})^2}. \quad (18)$$

EMIS is a measure of the overall performance, determined by the number of data objects that were misclassified. EMIS is calculated only when an algorithm correctly identifies the number of clusters. Then, the assigned cluster label of each object is compared with its actual cluster label. If any of them do not match, a misclassification has occurred

We compared performance of fzGASCE with eight genetic algorithm methods that also use FCM, specifically, PBMF, MPC, HPK, AGFCM, XB, FS, PC and ACVI [6-10]. An earlier version of fzGASCE, fzGAE which does not include the fuzzy SC and defuzzification methods, was also used. For each dataset, the number of clusters, c, was set to the known number of clusters. All algorithms were run using a population size of 24 and a maximum number of 100 generations. The fuzzy partition of each chromosome was generated using the FCM algorithm with 10 iterations, and the fuzzifier factor, m, was set to 1.25. We repeated the experiment 100 times and averaged the performance of each algorithm using values of COR, EVAR and EMIS.

ASET1 dataset

ASET1 contains five clusters in a 2-dimensional data space. The clusters are well-separated and of the same size. Performance of all algorithms is shown in Table 1. All algorithms had very low EMIS measures, indicating that they grouped the data points into the correct clusters. However, fzGASCE outperformed all algorithms by all three measures and fzGAE performed better than the other methods with the exception of fzGASCE. This comparison illustrates the advantage of using the method of Le et al. [12] in the fitness function. The use of fuzzy SC [13] in fzGASCE improved performance particularly in escaping local optima.

Table 1

Results with ASET1

Algorithm	COR	EVAR	EMIS
fzGASCE	**1.000**	**0.000**	**0.000**
fzGAE	0.640	0.500	0.000
PBMF	0.510	0.590	0.000
MPC	0.290	0.970	0.000
HPK	0.100	5.010	0.021
AGFCM	0.600	2.800	0.000
XB	0.490	1.450	0.000
FS	0.120	1.100	0.070
PC	0.230	1.040	0.000
ACVI	0.200	2.490	0.011

ASET2 and ASET3 datasets

Table 2

Results with ASET2

Algorithm	COR	EVAR	EMIS
fzGASCE	**1.000**	**0.000**	**0.000**
fzGAE	0.710	0.380	0.000
PBMF	0.600	0.450	0.000
MPC	0.610	0.860	0.000
HPK	0.120	5.240	0.000
AGFCM	0.650	1.490	0.000
XB	0.640	0.430	0.000
FS	0.520	0.840	0.011
PC	0.620	0.890	0.000
ACVI	0.100	2.100	0.000

Table 3

Results with ASET3

Algorithm	COR	EVAR	EMIS
fzGASCE	**1.000**	**0.000**	**0.000**
fzGAE	0.450	0.750	0.000
PBMF	0.340	1.000	0.000
MPC	0.420	0.820	0.002
HPK	0.010	1.910	0.037
AGFCM	0.340	2.380	0.000
XB	0.410	0.900	0.000
FS	0.400	0.880	0.000
PC	0.450	0.700	0.000
ACVI	0.170	4.650	0.000

The ASET2 contains five well-separated clusters in a 3-dimensional data space while the ASET3 contains 11 clusters in a 5-dimensional data space. Performance of all algorithms on ASET2 and ASET3 are shown in Tables 2 and 3 respectively. On both datasets, fzGASCE outperformed all other algorithms, while fzGAE was the second at best.

ASET4 dataset

This dataset is non-uniform with three clusters in a 2-dimensional data space (Figure 1). Table 4 shows the algorithm performance. HPK and AGFCM both failed to determine the number of clusters. fzGASCE not only outperformed the other algorithms, with COR=1.0, but also was the only algorithm that grouped all data points into the correct clusters. Although fzGAE performed better than the remaining algorithms, it failed to correctly group data points into clusters, similar to other methods that used the maximum membership degree for defuzzification.

Table 4

Algorithm performance on the ASET4 dataset

Algorithm	COR	EVAR	EMIS
fzGASCE	**1.000**	**0.000**	**0.000**
fzGAE	0.900	0.100	0.107
PBMF	0.700	0.300	0.107
MPC	0.050	0.960	0.107
HPK	0.000	5.770	-
AGFCM	0.000	8.470	-
XB	0.040	0.960	0.107
FS	0.020	3.480	0.107
PC	0.050	0.960	0.107
ACVI	0.080	0.920	0.107

IRIS dataset

The IRIS dataset contains three clusters corresponding to the three classes of Iris flowers [4]. The performance of the algorithms on this dataset is shown in Table 5. HPK and AGFCM again completely failed at detecting the number of clusters. fzGASCE outperformed other algorithms in detecting the number of clusters as well as in grouping data points into their own clusters.

Table 5

Algorithm performance on the IRIS dataset

Algorithm	COR	EVAR	EMIS
fzGASCE	**1.000**	**0.000**	**0.033**
fzGAE	0.880	0.120	0.040
PBMF	0.860	0.140	0.040
MPC	0.040	0.970	0.160
HPK	0.000	5.720	-
AGFCM	0.000	8.120	-
XB	0.050	1.010	0.040
FS	0.390	0.780	0.154
PC	0.080	0.920	0.115
ACVI	0.150	0.850	0.040

Wine dataset

Table 6

Algorithm performance on the WINE dataset

Algorithm	COR	EVAR	EMIS
fzGASCE	**1.000**	**0.000**	**0.213**
fzGAE	0.860	0.140	0.303
PBMF	0.000	2.050	-
MPC	0.000	2.810	-
HPK	0.000	6.760	-
AGFCM	0.000	9.210	-
XB	0.270	1.010	0.303
FS	0.000	5.720	-
PC	0.110	0.920	0.303
ACVI	0.090	0.910	0.303

The Wine dataset contains information on 13 attributes of three classes of wines [4]. Results on this dataset are shown in Table 6. Only fzGASCE and fzGAE identified the correct number of clusters, with COR values of 1 and 0.86, respectively. Among the other algorithms, only XB and ACVI detected the correct number of clusters but only with low COR values. Overall, fzGASCE outperformed all of the tested algorithms.

5 Conclusions

We have presented fzGASCE, a novel fuzzy clustering algorithm that combines the Genetic Algorithm with the fuzzy subtractive clustering and Bayesian based cluster evaluation methods. fzGASCE solves the problem of data clustering in the absence of information on the real number of clusters. fzGASCE outperformed other methods on both artificial and real datasets, and performed particularly well on datasets with clusters that differed in size, not only in predicting the correct number of clusters but also in grouping the data points into the correct clusters. fzGASCE is therefore appropriate for real-world problems, where the data densities are not uniformly distributed. In future work, we will develop fzGASCE into a more powerful tool for cluster analysis of microarray gene expression data.

6 References

[1] J.C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, NewYork, 1981.

[2] K. Loquin and O. Strauss, "Histogram density estimators based upon a fuzzy partition", Statistics and Probability Letters, Vol. 78, pp. 1863–1868, 2008.

[3] M.C. Florea, A.L. Jousselme, D. Grenier, and E. Bosse, "Approximation techniques for the transformation of fuzzy sets into random sets", Fuzzy Sets and Systems, Vol. 159, pp. 270–288, 2008.

[4] A. Frank and A. Asuncion, (2010) Machine Learning Repository. [Online]. http://archive.ics.uci.edu/ml.

[5] Y. Fukuyama and M. Sugeno, "A new method of choosing the number of clusters for the fuzzy c-means method", in: Proc. Fifth Fuzzy Systems Symp., 1989, pp. 247–250.

[6] A. Halder, S. Pramanik, and A. Kar, "Dynamic Image Segmentation using Fuzzy C-Means based Genetic Algorithm," International Journal of Computer Applications, Vol. 28, pp. 15-20, 2011.

[7] A. Ghosh, N. S. Mishra, and S. Ghosh, "Fuzzy clustering algorithms for unsupervised change detection in remote sensing images," Information Sciences, Vol. 181, pp. 699-715, 2011.

[8] Z. Lianjiang, Q. Shouning, and D. Tao, "Adaptive fuzzy clustering based on genetic algorithm," In Proc. of 2nd conference on advanced computer control, Shenyang China, 2010, pp. 79-82.

[9] T.C. Lin, H.C. Huang, B.Y. Liao, and J.S. Pan, "An Optimized Approach on Applying Genetic Algorithm to

Adaptive Cluster Validity Index," International Journal of Computer Sciences and Engineering Systems, Vol. 1, pp. 253-257, 2007.

[10] Y. Liu and Y. Zhang, "Optimizing Parameters of Fuzzy c-Means Clustering Algorithm," In Proc. of the 4th conference on Fuzzy Systems and Knowledge Discovery (FSKD '07), Vol. 1, 2007, pp. 633-638.

[11] M.K. Pakhira, S. Bandyopadhyay, and U. Maulik, "Validity index for crisp and fuzzy clusters", Pattern Recognition, Vol. 37, pp. 481–501, 2004.

[12] Thanh Le and Katheleen Gardiner, "A validation method for fuzzy clustering of gene expression data," Proc. Intl' Conf. on Bioinformatics and Computational, Las Vegas USA, 2011, Vol. 1, pp. 23-29.

[13] Thanh Le and Tom Altman, "A new initialization method for the Fuzzy C-Means Algorithm using Fuzzy Subtractive Clustering," Proc. Intl' Conf. on Information and Knowledge Engineering, Las Vegas USA, 2011, Vol. 1, pp. 144-150.

[14] X.L. Xie and G. Beni, "A validity measure for fuzzy clustering", IEEE Trans. Pattern Anal. Mach. Intell, Vol. 13, pp. 841–847, 1991.

[15] L. Xu and M.I. Jordan, "On convergence properties of the EM algorithm for Gaussian mixtures", Neural Computation, Vol. 8, pp. 129–151, 1996.

[16] M. Garey and D. Johnson, "Computers and Intractability- A Guide to the Theory of NP-completeness," Freeman, 1979.

Job Recommendation Systems for Enhancing E-recruitment Process

Shaha T. Al-Otaibi [1] **and Mourad Ykhlef** [2]
[1] College of Computer and Information Sciences, Princess Nora Bint Abdulrahman University, Riyadh, KSA
[2] College of Computer and Information Sciences, King Saud University, Riyadh, KSA

Abstract - *The Internet caused a substantial impact on the recruitment process through the creation of e- recruiting platforms that become a primary recruitment channel in most companies. While companies established job positions on these portals, job-seeker uses them to publish their profiles. E-recruitment platforms accomplished clear advantages for both recruiters and job-seekers by reducing the recruitment time and advertisement cost. However, these platforms suffer from an inappropriateness of traditional information retrieval techniques like the Boolean search methods that caused many applicants missed the opportunity of recruiting. Recommender system technology aims to help users in finding items that match their preferences; it has a successful usage in a wide-range of applications to deal with problems related to information overload efficiently. In order to improve the e-recruiting functionality, many recommender system approaches have been proposed. This paper will analyze e-recruiting process and related issues for building personalized recommender systems of candidates/job matching.*

Keywords: Recommender systems; Collaborative filtering; Content-based filtering; Hybrid approach; e-recruiting; Similarity measure.

1 Introduction

With the increasing volume of information available online, recommender systems have become a daily tool for Internet users, providing them with desirable help in finding information. The recommender systems used to determine the interested items for a specific user by employing a variety of information resources that related to users and items. In the mid-1990s, the term recommender system was published for the first time in information system literature. Recommender systems are being broadly accepted in various applications to suggest products, services, and information items to latent customers. Many e-commerce applications joint recommender systems in order to expand customer services, increase selling rates and decrease customers search time. For example, a wide range of companies such as the online book retailer Amazon.com [1], books [2], news articles [3]. Additionally, Microsoft provides users many recommendations such as the free download products, bug fixes and so forth [4]. All these companies have successfully set up commercial recommender systems and have increased web sales and improved customer fidelity.

For many years, information system supports in human resource management have been mainly restricted in storing and tracking applicants' data through the applicant management systems. These systems support the internal workflows and communication processes between the human resource management department and the other departments. Recently, the increased amount of digital information and the emergence of e-business reform the way companies conduct business in different aspects. Initially, simple solutions are applied such as posting the job ads on the career unit of the corporate website. Then, based on the experiences gained from these first implementations, the opportunities are realized establishing other changes and hence, implementing enhanced e-recruitment platforms.

The Internet-based online recruiting platform or e-recruitment platform is one of the most successful e-business changes, which changed the way companies employ candidates. These platforms spread in the recent years because the recruiting of the appropriate person is a challenge faced most companies, as well as the unavailability of certain candidates in some skill areas has long been identified as a major obstacle to companies success [5]. The online channels like Internet job portal, social media applications or a firm's career website have driven this development. While the companies established job positions on these portals, job-seeker uses them to publish their profiles. For each posted job, thousands of resumes received by companies. Consequently, a huge volume of job descriptions and candidate resumes are becoming available online. This vast volume of information gives a great opportunity for enhancing the matching quality; this potential is unused since search functionality in recruiting applications is mainly restricted to Boolean search method. The need increases for applying the recommender system

technologies that can help recruiters to handle this information efficiently [6] [7]. Many researches have been conducted to discuss different issues related to the recruiting problem as well as, the applying of recommender system technologies. However, job recommendation is still a challenging domain and a growing area of research. In order to support this research area, we conduct a comprehensive study for job recommendation. We will discuss the e-recruitment problem and present the different issues related to applying recommender systems in candidates/job matching.

This paper is organized as follows, in section 2 we present the recruiting process and its platforms. In section 3, we demonstrate the typical methods of various recommender system techniques. Section 4 illustrates the job recommendation related issues and presents a case study for applying recommender systems in candidates/job matching. Finally, we conclude this work in section 5.

2 The recruiting process

Recruiting process is a core function of human resource management treating the labor as one of the important factors of production [6]. The key objective of the recruiting process is to hire candidates who are valuable for the company [8]. Two viewpoints are distinguished: from recruiters and job seekers. The recruiters generate the job description by determining the set of requirements and constraints on skills, expertise levels, and degrees. The job-seeker, on the other hand, generates his/her CV by specifying the academic background, previous work experience and skills [9]. The IT support for the recruiting activities is ranging from attracting and finding talent to choose and retain candidates [10]. The degree of process integration represents the complexity of using e-recruitment solutions [11].

Färber et al. [6] demonstrated in their proposed model the relationship between recruiting tasks and divided the recruiting process into two main phases: The attraction phase and the selection phase, both phases contain a planning and an execution part. The planning part determines the overall strategy and actual measures to attract valuable employees as well as, the explicit selection methods. The execution part comprises the employer branding activities that include all long-term marketing measures that attracting qualified candidates. The attraction phase aims to generate a description for open job positions. The selection phase starts with the pre-screening of resumes and other submitted materials. Then, the final selection of candidates is conducted by comparing the remaining set of candidates that has not been filtered out in the screening phase. Finally, the applicant management serves as a secondary function; it consists of the contact of applicants, the management of applicant data and associated processes such as directing applications to organization's members that involved in the selection decision. Figure1 represents the recruiting process that adapted from [12]. Additionally, Carroll et al. [13] presented four phases of the recruiting process: an assessment of job position that needs to be filled, a description job profile, the construction of a job description and a candidate specification. Moreover, Breaugh and Starke composed the recruiting process into five main tasks: short-term and long-term candidate attraction, applicant management, pre-selection as well as the final selection of candidates. Short-term and long-term marketing measures are establishing the attractive employer image that intended to attract qualified candidates [14].

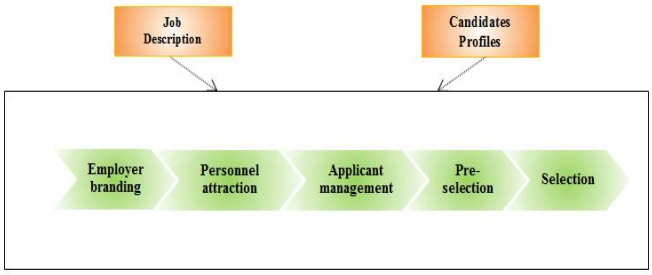


Figure 1: Recruiting process

2.1 E-recruitment platforms

The e-recruitment is a system for quickly reaching a large set of potential job-seekers. E-recruiting has attractive growth since the late 1990s when the rapid economy changes produced a high demands for qualified candidates that the labor market could not fully satisfy. The e-recruiting platforms such as corporate homepages and job portals (e.g. monster.com) have driven this development. The International Association of Employment websites[1] mention that there are more than 40,000 employment sites helping job-seekers and recruiters worldwide [9]. While companies send open job positions on these portals, job-seekers use them to publish their profiles, this caused a vast amount of job descriptions and candidates' profiles are becoming available online. However, the adoption of these e-recruiting platforms accomplishing cost savings, effectiveness, and suitability for both recruiters and job-seekers [15]. Many online recruiting platforms suffer from an inappropriateness of Boolean search methods for matching applicants with job requirements. Consequently, a large number of candidates missed the opportunity of recruiting [12]. Actual practices and theoretical thoughts show that this search type is insufficient for achieving a good fit between candidate aptitudes and job requirements [6]. Researchers have identified different reasons why organizations implement e-recruiting platforms; they discussed several challenges that faced the organizations when implementing IT support for

[1] http://www.employmentwebsites.org/

their recruiting activities. Lang et al. [12] presented detailed information about drivers, challenges and consequences of e-recruiting platforms.

2.2 Categories of e-recruitment platforms

In order to give the reader a better understanding of the e-recruiting platforms, we present the six categories of e-recruiting sources that presented by Lee [15]: (1) General-purpose job boards (e.g. Monster.com; HotJobs.com) that provide complete online recruiting functions. While job-seekers search jobs by category such as experience, location, education or any combination of these attributes, recruiters search applicants databases by skills, experience, preference, education, salary or any combination of keywords. (2) Niche job boards (e.g. Dice.com; Erexchange.com) serve the specialized markets such as a particular occupation, industry, education or any combination of specialties. (3) E-recruiting application service providers (e.g. RecruitUSA; PeopleClick) present a collection of services such as recruitment software, recruitment process management, education and training. (4) Hybrid recruiting service providers (e.g. magazines and Journals) are the traditional means that provide e-recruiting services. (5) E-recruiting consortium (e.g. DirectEmployers.com; NACElink.com) is a search engine drives traffic directly to a member's career website. (6) Corporate career website is an employment source most commonly used by Fortune 500 companies where the use of the corporate career website is a regular extension of e-business applications.

3 Background of recommender systems

The recommender system approaches are classified into the following main four categories: Collaborative filtering, Contend-based filtering, Knowledge-based and Hybrid approaches [16]. The descriptions of different techniques are presented in the following paragraphs.

(1) Collaborative Filtering (CF) is one of the most successful approaches for building recommender systems. It applies the known preferences of a set of users to predicate the unknown preferences for new users. The fundamental assumption of CF is that if users x and y rate n items similarly, or have similar behaviors. Hence, they will rate other items similarly [17]. CF approaches have the capability of working in domains where items contents are difficult to obtain or cannot be parsed automatically.

(2) Content-Based Filtering (CBF) is treated as information retrieval problem or machine learning problem. In information retrieval problem, the document representations have to be matched to user representations on textual similarity while, in machine learning problem, the textual content of the representations are combined as feature vectors, which are used for training a prediction algorithm [16]. The CBF recommends items whose content is similar to the content that the user has previously viewed or selected [2]. There are two main tasks related to CBF recommender systems, the User profiling and the Item representation. User profiling is one of most challenging tasks in CBF recommender systems that deal with acquiring, extracting and representing the features of users. User's profile may contain different types of information such as the selected items, ratings of items, and user's demographic data, etc. [18]. Item representation is also an important issue in CBF recommender systems. Items can be a structural data represented by the same set of attributes, and there are specific values that the attributes may have. Several approaches for learning a structural data used such as machine learning techniques. Additionally, unstructured data may occur in some applications such as unrestricted texts in news articles. In this type, there are no attribute names with well-defined values. A common approach to deal with free text fields is to exchange the text to a structured representation [19].

(3) Knowledge-Based Approach, this type of recommender systems attempts to suggest objects based on inferences about user's needs and preferences [20]. This approach assists users in the determination of suitable solutions from complex product and service assortments. These solutions based on exploiting deep knowledge about the product domain to figure out the best wishes of the customer [21]. They can use rules and patterns to recommend items based on functional knowledge of how a specific item meets a particular user need [20].

(4) Hybrid Approach, all recommendation approaches that mentioned above have characteristics and challenges. To get better performance and overcome challenges, these approaches have been combined. In general, collaborative filtering is integrated with other techniques in an attempt to avoid these challenges [20]. Authors of [20] and [22] presented different ways to integrate collaborative filtering, content-based filtering and knowledge-based approaches into a hybrid recommender system.

4 Job recommendation systems

Recent researches show that the increasing demands of Information System technologies for human resource management in general and recruiting processes in particular. Most companies put the focus on their own e-recruiting platforms as primary recruitment channels. Job ads are published automatically on the job portal as soon as they are entered into the system. On the other hand, the applicant creates a profile to apply it for one of the listed job positions. The user profile is stored in the system, letting the applicant reuse it for other job position. The last functionality gives the companies possibility to create the applicants pool. Thus, the companies achieved a uniform view for all applicants' data in one candidate pool. This pool is used by the recruitment department to find the applicant documents.

Appropriate applicants' documents are directed to the human resource departments for more processing. In addition, the system supports all required communication processes as well as tracks applicant status inside the application process [11].

As mentioned previously, the e-recruiting platforms are usually based on Boolean search and filtering techniques that cannot sufficiently capture the complexity of a person-job fit as selection decisions [23]. Many literatures have been applied the recommender system concept into the job problem. Malinowski et al. [24] determined that, we must consider unary attributes such as individual skills, mental abilities and personality that control the fit between the individual and the tasks to be accomplished, as well as the relational attributes that determine the fit between the individual and the upcoming team members. In this context literatures usually distinguish between (1) person-job, (2) person-team and (3) person-organization fits [25]. Thus, the recruitment approach must cover all these aspects. Keim argues that transferring recommender system approach to search for persons is a challenging but promising goal [26]. Therefore, many recommendation approaches applied for matching candidates and jobs to overcome the previous challenges of holistic e-recruiting platforms [8].

4.1 System requirements for candidates/job recommendation

There are major requirements presented in literatures that should be derived when recommending candidates for a specific job [23] [26] [24].

1. The matching of individuals to job depends on skills and abilities that individuals should have.
2. Recommending people is a bidirectional process that needs to take into account the preferences not only of the recruiter but also of the candidate.
3. Recommendations should be based on the candidate attributes, as well as the relational aspects that determine the fit between the person and the team members with whom the person will be collaborated.
4. Individual is considered to be unique; we cannot choose a single person several times such as a movie or book.

Job recommendation problem is bidirectional recommendation between job seeker and job. The recommendation process can be divided into two parts: job recommendation and job-seeker recommendation. The design idea of these two parts is the same roughly [27] [23]. For a job-seeker, the job with higher matching degree should be recommended to him. Similarly, for a job, the job-seeker with higher matching degree should be recommended to it [27]. In general, the ranking items either are the top n candidates that best fit the job in consideration or the top n job profiles that best fit the candidates' preferences. Additionally, Fazel-Zarandi and Fox mentioned that skills requirements matching need to distinguish between must-have and nice-to-have requirements in the matching process. Must-have requirements are constraints that should be possessed by the applicant, whereas nice-to-have requirements are preferences that are taken into consideration when ranking applicants [9]. Figure 2 summarizes the job recommendation requirements in a unified model.

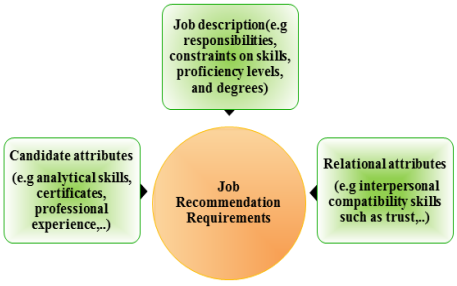


Figure 2: Model of system requirements for candidates/job recommendation

4.2 Job recommendation information

Candidates and jobs should be matched based on certain criteria that used as indicators of performance on the job. In selection theory, the available information at a certain time of the decision selection is called predictor data which comprises the individual attributes. The actual selection method is called predictor. The prediction process is referred to the assessment of the criteria using the predictor data and a method-specific way of data combination [6]. However, to construct candidate profiles, the meta-data extracted from existing resumes. Rafter and Smyth proposed a system that builds user profile in recruitment environment directly from analyzing the behaviors of web users. In this system, user profiles are constructed by passively detecting the click-stream and read-time behavior of users [28]. Malinowski et al. used an input data for their CV-recommender: demographic data, educational data, job experience, language skills and IT skills, awards, publications, others [23]. In general, candidate's profile is composed of three sections.

1. Personal information about the employee, such as the first name, last name, and location.
2. Information about the current and past professional positions held by the candidate. This section may contain company names, positions, company descriptions, job start dates, and job finish dates. The company description field may further contain information about the company (e.g., the number of employees and industry).
3. Information about educational experiences, such as university names, degrees, fields of education, start and finish dates [29].

Additionally, for collaboration measures, candidate may be asked to rate the job profiles using 5 point scale ranging from 1 to 5. Candidates were asked to evaluate whether the profiles interested to them with respect to their career perspectives and planning [23]. From these meta-data, a number of

features can be extracted to train and test recommendation [29]. On the other hand, the job profile should be constructed to describe the requirements and listing of all relevant skills that an employee for this job should have [8]. Moreover, the quality of the recommendation system can be assessed using statistical accuracy metrics such as the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) or Correlation calculations [30] [23] [17].

4.3 Job recommendation architecture

Laumer and Eckhardt [8] proposed system architecture that aligns recommender systems with the recruiting process based on the preceding holistic e-recruiting architecture provided by Lee [15]. They added new processes that supporting the development of job profiles and automated recommendation approaches. In his architecture proposal, Lee presented a workflow management subsystem linked to a database management subsystem as the central component. All information related to recruiting activities is stored in the database. Any subsystem can have access to data stored by another subsystem and processes can include other processes or execute them. The integrated architecture for employee recruitment and recommender systems is built on the workflow management subsystem and database to manage the information flow and storage. For the integration of recommender systems, they added two important parts: First, a process to build job profiles that describing the job requirements and listing all related skills an employee for this profile should have. Second, they integrated a person-job recommender in the recruitment process as a process step in the selection phase. Finally, matching candidate and jobs can be managed by automated recommendation approaches [8]. Figure 3 illustrates the integrated system architecture for job recommendation.

Figure 3: The integrated architecture for job recommender system

4.4 Case study: an example of recommending candidates for specific job

In order to understand the job recommendation problem, we present a simple and concrete example for matching candidate with job requirements. We focus on measurable skills possessed by human resources. This example applies a content-based recommendation approach that used the attributes related to both job and candidates. As mentioned before in content-based approach in section 3, we must construct a profile for each item, which is a record representing the important features of that item. In job case, the candidate's profile consists of some features that required for a specific job. Similarly, the job's profile consists of the job requirements that should be possess by candidates. For simplicity, we consider only few features that might be relevant to a recommendation system. The task of a job recommender system is to retrieve a list of candidates' CVs for a new job position. We conduct this example using one job description and list of 5 prospective candidates CVs. The job description was downloaded from Careers portal website[2]:

- Job title: Computer System Administrator.
- Job description: the prospective employee will monitor, operate and supervise the internal computer systems of an organization.

[2] www.careersportal.ie

• Qualifications required: BSc certificate in Software Engineering, Computer Programming or IT and four years of experience in IT sector, especially as Systems Analyst or System Programmer.
• Skills: English language skill (1-low, 2-medium, 3-excellent) and Oracle developer skill.

The candidates' CVs were downloaded from BSR site[3]. As mentioned above, the first step to determine the best fit between candidates and job description is building the job profile and the prospective employees' profiles. We extract some features from employee resumes and job description to build both profiles. Then, we estimate the model parameters by creating a rating matrix $R_{x,y}$, where x represents the job and y represents the candidate CVs.

$$R_{x,y} \begin{cases} 1 \text{ (TRUE =”Exist”)} & \text{‘If the target attribue is existed} \\ 0 \text{ (FALSE =“not Exist”)} & \text{‘If the target attribute isn't existed} \\ \text{“Value”} & \text{‘For quantity attributes} \end{cases}$$

The rating matrix $R_{x,y}$ transformed by treating the values of candidate's attributes as ratings of all the attributes extracted from the resumes using any similarity measures. That means the job profile as well as the candidates' profiles represented as vectors. We applied three measures in this example: Cosine Similarity, Euclidean Distance [31] and New Jaccard Measure [32].

The profiles vectors are constructed as the following: 0 (MSc not required), 1(BSc required), 1(if one of these majors: Software Engineering, Computer Programming or IT), 1(if he/she worked in IT sector), 1(if the candidate's experience more than 4 years), 1-3 (for English skill levels), 1(if the candidate has Oracle developer skill).

The resultant job's vector is [0 1 1 1 1 3 1] and the resultant candidates' vectors are: 1st person: [0 0 1 0 0 2 1], 2nd person: [0 1 0 1 1 3 0], 3rd person: [0 1 1 1 1 2 1], 4th person: [0 1 0 1 1 2 0], and 5th person: [1 1 0 1 1 1 0]. The candidates' ranking after applying the Cosine Similarity, Euclidean Distance and new Jaccard Measure is presented in table 1.

Table 1: Ranking of candidates for the job position using three similarity measures.

Cosine Similarity	Euclidean Distance	New Jaccard Measure
3rd person 0.99	3rd person 1.0	3rd person 0.94
2nd person 0.82	2nd person1.41	2nd person 0.83
4th person 0.81	4th person 1.7	4th person 0.78
1st person 0.70	1st person 2.0	1st person 0.61
5th person 0.67	5th person 2.65	5th person 0.48

[3] www.bestsampleresume

This example aims to find a candidate who fits best to the requirements of job profile [0 1 1 1 1 3 1]. Based on the three similarity measures, 3rd person is the best candidate who fits job requirements, followed by 2nd person and 4th person. The 1st person and 5th person are the least appropriate candidates for the job requirements.

5 Conclusion and future works

In this paper, we used a literature analysis of many journals and proceedings related to the recruiting process and the job recommendation researches. We have seen from our literature review and from the challenges that faced the holistic e-recruiting platforms, an increased need for enhancing the quality of candidates/job matching. The recommender system technologies accomplished significant success in a broad range of applications and potentially a powerful searching and recommending techniques. Consequently, there is a great opportunity for applying these technologies in recruitment environment to improve the matching quality. This paper analyzed the e-recruiting process and the different aspects related to applying the recommender systems in candidates/job matching problem. Additionally, in order to give a clear understanding for job recommendation problem, a case study of applying three measures for matching candidates with job position was presented. Finally, we plan as a continuation of this work to present a survey of job recommendation approaches that have been proposed to produce the best fit between jobs and candidates. We will introduce state of the art of job recommendation as well as, a comparative study for its approaches.

6 References

[1] G. Linden, B. Smith and J. York, "Amazon.com Recommendations: Item-to-Item Collaborative Filtering," Published by the IEEE Computer Society, IEEE Internet Computing, vol. [7] , no. 1, pp. 76-80, 2003.

[2] R. J. Mooney and L. Roy, "Content-Based Book Recommending Using Learning for Text Categorization," in In Proceedings of DL '00: Proceedings of the Fifth ACM Conference on Digital Libraries, New York, NY, 2000.

[3] A. Das, M. Datar, A. Garg and S. Rajaram, "Google News Personalization Scalable Online Collaborative Filtering," in In Proceedings of the 16th International Conference on World Wide Web, WWW '07, Banff, Alberta, CANADA, 2007.

[4] G. Shani and A. Gunawardana, "Evaluating Recommendation Systems," Springer, pp. 257–298, 2011.

[5] S. Laumer and A. Eckhardt, "Analyzing It Personnel's Perception of Job-Related Factors in Good and Bad Times," in In Proceedings of the 2010 Special Interest Group on Management Information System's 48th annual conference on Computer

personnel research on Computer personnel research, Vancouver, BC, Canada, 2010.
[6] F. Färber, T. Weitzel and T. Keim, "An Automated Recommendation Approach to Selection in Personnel Recruitment," in In Proceedings of AMCIS, Tampa, FL, USA, 2003.
[7] X. Yi, J. Allan and W. B. Croft, "Matching Resumes and Jobs Based on Relevance Models.," in In Proceedings of SIGIR, New York, NY, USA, 2007.
[8] S. Laumer and A. Eckhardt, "Help to Find the Needle in a Haystack: Integrating Recommender Systems in an It Supported Staff Recruitment System," in In proceedings of the special interest group on management information system's 47th annual conference on Computer personnel research, Limerick, Ireland, 2009.
[9] M. Fazel-Zarandi and M. S. Fox, "Semantic Matchmaking for Job Recruitment: An Ontolgy Based Hybrid Approach.," in In Proceedings of the 3rd International Workshop on Service Matchmaking and Resource Retrieval in the Semantic Web at the 8th International Semantic Web Conference, Washington D.C., USA, 2010.
[10] S. Laumer, A. Eckhardt and T. Weitzel, "Electronic Human Resources Management In An E-business Environment," Journal of Electronic Commerce Research, vol. [11], no. 4, pp. 240-250, 2010.
[11] J. Malinowski, T. Keim and T. Wietzel, "Analyzing the Impact of IS Support on Recruitment Processes: An E-recruitment Phase Model," in In Proceedings of the ninth Pacific Asia conference on information systems (PACIS-2005), Bangkok, Thailand, 2005.
[12] S. Lang, S. Laumer, C. Maier and A. Eckhardt, "Drivers, Challenges and Consequences of E-Recruiting – A Literature Review," in SIGMIS-CPR'11, San Antonio, Texas, USA, 2011.
[13] M. Carroll, M. Marchington, J. Earnshaw and S. Taylor, "Recruitment in Small Firms: Processes, Methods and Problems," Employee Relations, vol. [21], no. 3, pp. 236-250, 1999.
[14] J. A. Breaugh and M. Starke, "Research on Employee Recruitment: So Many Studies, So Many Remaining Questions," Journal of Management, vol. [26], no. 3, pp. 405-434, 2000.
[15] I. Lee, "An Architecture for a Next-Generation Holistic E-Recruiting System," Communications of the ACM, vol. [50], no. 7, pp. 81-85, 2007.
[16] K. Wei, J. Huang and S. Fu, "A Survey of E-Commerce Recommender Systems," in In Proceedings of the International Conference on Service Systems and Service Management, USA, 2007.
[17] X. Su and T. M. Khoshgoftaar, "A Survey of Collaborative Filtering Techniques," Advances in Artificial Intelligence, pp. 421-425, 2009.
[18] A. Felfernig, M. Schubert and M. Mandl, "Recommendation and Decision Technologies For Requirements Engineering," in RSSE '10, Cape Town, South Africa, 2010.
[19] M. J. Pazzani and D. Billsus, "Content-Based Recommendation Systems," The Adaptive Web: Methods and Strategies of Web Personalization, vol. [4321], pp. 325–341, 2007.
[20] R. Burke, "Hybrid Recommender Systems:Survey and Experiments," User Model. User-Adapt. Interact, vol. [12], no. 4, pp. 331–370, 2002.
[21] A. Felfernig, "Koba4MS: selling complex products and services using knowledge-based recommender technologies," in In Proceedings of 7th IEEE International Conference on E-Commerce Technology, Munich, Germany, 2005.
[22] R. Burke, "Hybrid Web Recommender Systems," The Adaptive Web: Methods and Strategies of Web Personalization, vol. [4321], pp. 377-407, 2007.
[23] J. Malinowski, T. Keim, O. Wendt and T. Weitzel, "Matching People and Jobs: A Bilateral Recommendation Approach," in In Proceedings of the 39th Annual Hawaii International Conference on System Sciences, Hawaii, USA, 2006.
[24] J. Malinowski, T. Weitzel and T. Keim, "Decision Support for Team Staffing: An Automated Relational Recommendation Approach," Decision Support Systems, vol. [45], no. 3, pp. 429-447, 2008.
[25] T. Sekiguchi, "Person-Organization Fit and Person-Job Fit in Employee Selection: A Review of The Literature," Osaka Keidai Ronshu,, vol. [54], no. 6, pp. 179-196, 2004.
[26] T. Keim, "Extending the Applicability of Recommender Systems:A Multilayer Framework for Matching Human Resources," in In Proceedings of the 40th Annual Hawaii International Conference on System Sciences (HICSS'07), Hawaii, USA, 2007.
[27] H. Yu, C. Liu and F. Zhang, "Reciprocal Recommendation Algorithm for the Field of Recruitment," Journal of Information & Computational Science, vol. [8], no. 16, pp. 4061–4068, 2011.
[28] R. Rafter and B. Smyth, "Passive Profiling from Server Logs in an Online Recruitment Environment," in In Proceedings of the IJCAI Workshop on Intelligent Techniques for Web Personalization (ITWP 2001), Seattle, Washington, USA, 2001.
[29] I. Paparrizos, B. B. Cambazoglu and A. Gionis, "Machine Learned Job Recommendation," in In proceedings of the fifth ACM conference on Recommender systems,RecSys'11, Chicago, Illinois, USA, 2011.
[30] J. L. Herlocker, J. A. Konstan, A. Borchers and J. Riedl, "An Algorithmic Framework for Performing Collaborative Filtering," in In Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA, 1999.
[31] A. Rajaraman, WalmartLabs and J. D. Ullman, "Ch9: Recommendation Systems," in Mining of Massive Datasets, California, USA, Stanford University, 2011, pp. 287-321.
[32] A. Belkhirat, A. Belkhir and A. Bouras, "A New Similarity Measure for the Profiles Management," in UKSim 13th International Conference on Modelling and Simulation, Cambridge, United Kingdom, 2011.
[33] R. Burke, "Integrating Knowledge-Based and Collaborative-Filtering Recommender Systems," in In Proceedings of the AAAI Workshop on AI in Electronic Commerce, Orlando, Florida, USA, 1999.
[34] D. H. Lee and P. Brusilovsky, "Fighting Information Overflow with Personalized Comprehensive Information Access: A Proactive Job Recommender," in In Proceedings of the 3rd Conference on Autonomic & Autonomous System, Athens, Greece, 2007.
[35] G. Adomavicius and A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," IEEE Transactions on Knowledge and Data Engineering, vol. [17], no. 6, pp. 734-749, 2005.
[36] R. Burke, "Knowledge-Based Recommender Systems," Encyclopedia of Library and Information Systems, vol. [69], no. 32, 2000.
[37] U. Hanani, B. Shapira and P. Shoval, "Information Filtering: Overview of Issues and Systems," User Modeling and User-Adapted Interaction, vol. [11], pp. 203–259, 2001.

An Information System Modeling with UML2 for Reports Archiving for the Knowledge Management of a School Structure

S. Demigha

CRI (Centre de Recherche en Informatique), Université de Paris 1 Panthéon-Sorbonne, Paris, France

Abstract - *This paper describes the modeling of an Information System (IS) with UML2 of a school structure in order to implement a real Information System in the UFEC Institute in Paris (Graduate School). It presents requirements engineering able to develop the Information System. To design and implement this IS, we need to determine the different factors influencing this type of computer application. These factors determine the way data will be stored and retrieved. We use the last version of UML (Unified Modeling Language) to formalize and structure data. The UML is suited with the Object-Oriented approach and deal with a high level of abstraction of data particularly when it is about managing various and complex data. A school structure is splitting in many departments which necessitate a best organization of data and knowledge. This approach corresponds to our purposes and needs.*

Keywords: Information System, Modeling, UML2, Knowledge, Management.

1 Introduction

According to the literature [1],[2],[3],[4], an Information System (IS) is defined as a collection of hardware, software, data, people and procedures that work together to produce quality information. Silver and al, [5], defined Information Systems as follows: Information Systems are implemented within an organization for the purpose of improving the effectiveness and efficiency of that organization. Capabilities of the Information System and characteristics of the organization, its work systems, its people, and its development and implementation methodologies together determine the extent to which that purpose is achieved.

Information Systems research is generally interdisciplinary concerned with the study of the effects of Information Systems on the behavior of individuals, groups, and organizations, [6], [7]. Hevner and al, [8], categorized research in IS into two scientific paradigms including behavioral science which is to develop and verify theories that explain or predict human or organizational behavior and design science which extends the boundaries of human and organizational capabilities by creating new and innovative artifacts.

Salvatore March and Gerald Smith, [9], proposed a framework for researching different aspects of Information Technology including outputs of the research (research outputs) and activities to carry out this research (research activities). They identified research outputs as follows:

- Constructs which are concepts that form the vocabulary of a domain. They constitute a conceptualization to describe problems and specify their solutions.
- A model which is a set of propositions or statements expressing relationships among constructs.
- A method to perform a task. Methods are based on a set of underlying constructs and a representation (model) of the solution space.
- An instantiation which is the realization of an artifact in its environment.

An Information System is not only the technology an organization uses, but also the way in which the organizations interact with the technology and the way in which the technology works with the organization's business processes. Information Systems (IS) are distinct from Information Technology (IT) in that an Information System has an Information Technology component that interacts with the processes components, [10].

Information Technology departments in larger organizations tend to strongly influence Information Technology development, use, and application in the organizations, which may be a business or corporation. A series of methodologies and processes can be used in order to develop and use an Information System. Many developers have turned and used a more engineering approach such as the System Development Life Cycle which is a systematic procedure of developing an Information System through stages that occur in sequence.

A school structure can be considered as a large organization. This organization is splitting in various departments which take an important part for the organization. To organize these departments we need to develop a real Information System including all actors participating at this organization and gathering them in a single and common system.

For developing this system we have focused on the best and innovative methods, tools and techniques found in the literature based on high and efficient technologies and

adopted them to build our own IS with new purposes and propositions.

We choose a traditional cycle for developing this system by following a systematic procedure and retain the Object-Oriented, [11], approach with UML2 (Unified Modeling Language), [12], for the system modeling.

The object model allows for a higher level of abstraction when representing concepts from real word. It may represent all the complexity of a school structure without losing information and coherence. Moreover, it allows for an easy creation of new data types and a flexible scheme adapting to new emerging knowledge and the modifications of the model.

The paper is organized as follows:

- Section 2 presents the UFEC graduate school structure;
- Section 3 presents the users requirements of the IS;
- Section 4 describes the object model retained to the school structure;
- Section 5 is the conclusion with further research works in progress;

Section 2 presents the overall architecture of the UFEC school.

2 The UFEC graduate school structure

The UFEC School is splitting in 8 departments:

1. The Supervisor Department;
2. The Training Department;
3. The Administrative Department;
4. The Human Resources Department;
5. The Communication Department;
6. The Placement Department;
7. The e-learning and virtual University Department;
8. And the Security Department.

Figure 1 presents the global organization chart of the UFEC school.

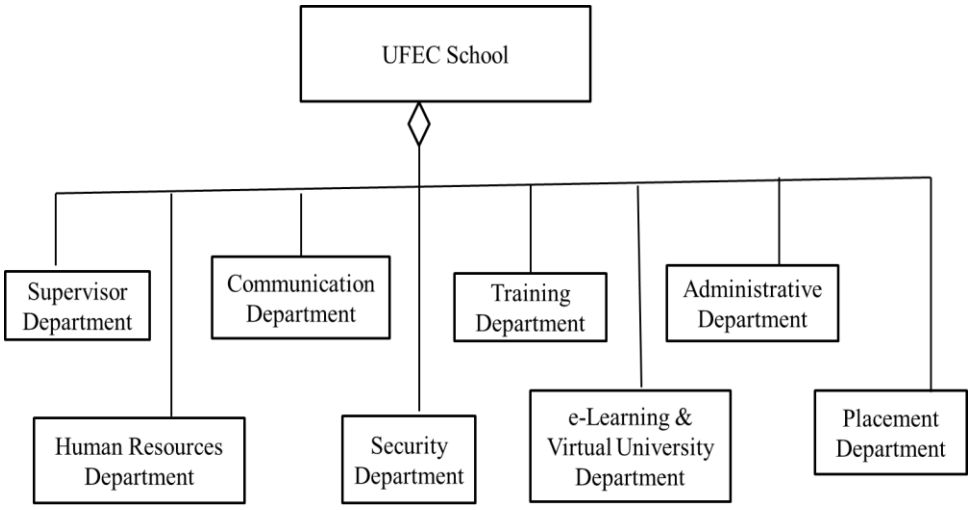


Figure 1: The general organization chart of the UFEC School

Section 3 presents the users requirements of the IS.

3 Users requirements

The development of the IS requires to set up specifications from user requirements. We have defined these specifications from the management of the UFEC school where the Information System will be implemented. We use UML to express these requirements. The UML (Unified Modeling Language) is a modeling language and has not a notion of development process, which must accompany a method. The dictionary defines a method as a systematic or orderly procedure, [13].

3.1 Staff using the IS

The public participating at the school organization belong to eight categories:

- *Supervisor staff:* aim to supervise the school.
- *Training staff:* aim to organize pedagogical contents and follow-up of students.
- *Lecturer staff:* aim to assure lessons, lead scientific works of students and evaluate them.
- *Student staff*: are concerned by the training and learning stages (Master and PhD).
- *Administrative staff:* are concerned by the administrative people (Internal and External).
- *Human Resources staff*: are advocated for both the school and the people who work in the company.
- *Communication staff:* people who disseminate information between the departments of the school and outside the school (enterprises,…).
- *Security staff:* people who control access to data; access to data must be controlled.

  Access to the system is granted either to the location of the user: outside the school (enterprises or external people) and inside the school (employees and students).

3.2 Characteristics

Data have been collected through the eight categories that we have mentioned in the previous section. These data must be analysed with UML2 and will be described in details in the following section.
Section 4 describes the Object-Oriented model retained to the school structure.

4 The object modeling with UML2

Through a model we aim to provide a better understanding of the system under development. UML2 model permits an application's design to be evaluated and critiqued before implementation, [13].
Changes are easy and less expensive to make when they are made in the early phases of the software lifecycle. Models help us capture and record our software design decisions as we progress toward an implementation. The UML2 defines a diagrammatic notation for describing the artifacts of an Object-Oriented analysis and design. We can visualize, specify, construct and document our software application. As our IS become ever larger and ever more complex we need to manage that complexity and, in a sense, simplify it which allows us to have a better understanding of it.
Finally, from our UML2 diagrams we can derive programming language code. This is referred to as forward engineering – the generation of code from UML2 models. This is an approach we advocate through this paper. The models are the core of our designs. In this paper we only focused on the modeling and the code for implementation will be programmed later in another paper. The object modeling is based on three models: the functional model, the object model and the dynamical model.

4.1 The functional model

It represents the functionality of the system and allows the modeling of users' expectations. There are two basic concepts in functional modeling: 'Users' that utilize the system, and 'Use Cases' that represent the utilization of the system by the users.

4.1.1 Users

Users are actors using the system. There are 5 types of users:

- *The lecturer:* accesses to anonymous data. Modifies only pedagogical data.
- *The student:* may access to certain type of data but has not to modify rights.
- *The supervisor:* accesses to all type of data. Can record, modify or create data.
- *The administrative employee:* accesses to administrative data. Can record, modify or create new data.
- *The system administrator:* has both rights of the supervisor and rights to set security levels for other users and create or suppress a user.

4.1.2 Use Cases

In this paper, we won't to describe all 'Use Cases' useful of our IS because of the complexity of the system and the number of pages that it may include.

Use Case "system Login/Logout"
- The user enters the system with a 'login' and a 'password' after identification
- **If** the system identifies the user as authorized **Then**
  It allows access
  Otherwise
  It denies access
  It locked at the user identification step
- A user must logout from the system
  - **If** a user is not using the system during a period **Then** The system locks the access
  End if

End if

Use Case "see a record"
- The user selects the option "see" a record
- The user enters criteria to find the record to read
- **If** the record exists **Then**
- The system displays it
- **Otherwise**
- It informs the user
- It waits for another transaction
- **End if**

Use Case "create a new record"
- The user selects the option "new" a record
- The user enters criteria to create the new record
- **If** there is no similar record in the system **Then**
- The record is created by the system
- **Otherwise**
- It informs the user
- It waits for another transaction
- **End if**

Use case "modify an existing record"
- The user selects the option "delete a record"
- **If** the user has the authorization **Then**
- The record may be deleted
- **Otherwise**
- The user is informed
  The system waits for another transaction

End if

Figure 2 illustrates the functional model of the UFEC School.

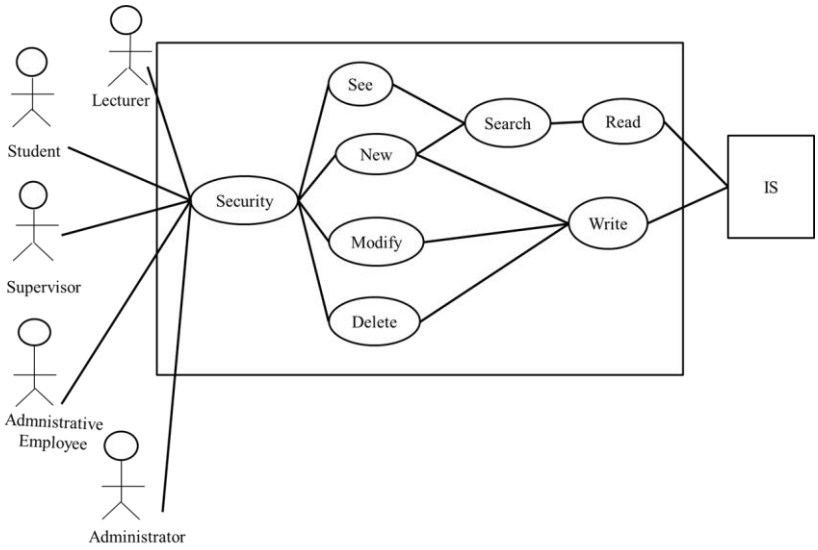


Figure 2: The functional model

4.2 The object model

The Object model (also called the Static model) describes the objects (concepts from the real world) and their relationships: description of the structure and characteristics of the objects (the "Classes"), and description of the various associations between the different objects (the "Associations").

4.2.1 Identification of Classes

Objects are characterized by having both 'state' and 'behavior'. The 'state' of an object is the information an object has about itself. For example, 'a student' may have *a name, a date of birth and a university matriculation number*. The 'behavior' of an object describes the actions the object is prepared to engage in. For example, we might ask 'a student object' for its *age*. This would involve the 'student object' performing a calculation based on *its date of birth* and *today's date*. The behavior of an object is described by the set of operations it is prepared to perform.
A 'graduate school' or a 'university' would typically have a large number of students. Unlike, real students, all 'student objects' exhibit the same behavior and carry the same knowledge about them-selves. We might model a 'student object' with *a name, date of birth* and *matriculation number*. The actual state values for 'two student objects' are different since 'university' or 'graduate school matriculation numbers' are unique. With a large 'university' or 'graduate school' population, we might, however, expect two or more students with the same *name* or two or more with the same *date of birth.*
How can we resolve this problem? All of our 'student objects' support a single abstraction that we may choose to call 'student'. We refer to the abstraction as the Class of the object.
The Class describes any number of 'student objects'. The 'Supervisor Class' describes any number of 'supervisor objects'. The Class describes the information and object holds to represent its state. The items of information are called *attributes (or properties).* The Class also defines the behavior of such objects, listing the operations they can perform, i.e. the messages they can receive. The effect of these operations is described by its *method.*

4.2.2 Identification of relationships

Objects enter into relationships with each other. One of the most important relationships is "Association". In general, "Association" should be used where two objects are not conceptually related but within the context of the problem need to make use of each other's services. For example, an interaction in which a single 'Administrative Employee' object is employed by a single 'Graduate School' object. They are associates in the sense that the 'Graduate School' object adopts the role of the 'Employer' while the 'Employee' object adopts the role of the 'Employee'. The 'Employee' could request the name of the 'Employer' while the 'Employer' could request the job title or salary of the 'Employee'.

4.2.3 The Class diagram

The Class Diagram lies in the fact its content delivers the primary elements in our program code, namely the Java Classes. A Class Diagram describes the types of objects in the system and the relationships that exist between them. A Class Diagram is an abstraction for all the possible object diagrams we might construct. A Class is also documented with its set of attributes and operations. The attributes represent the set of values each instance maintains the object's state. The set of operations are the messages an object of the Class may receive.

Figure 3 illustrates an extract of the Class Diagram of our IS.

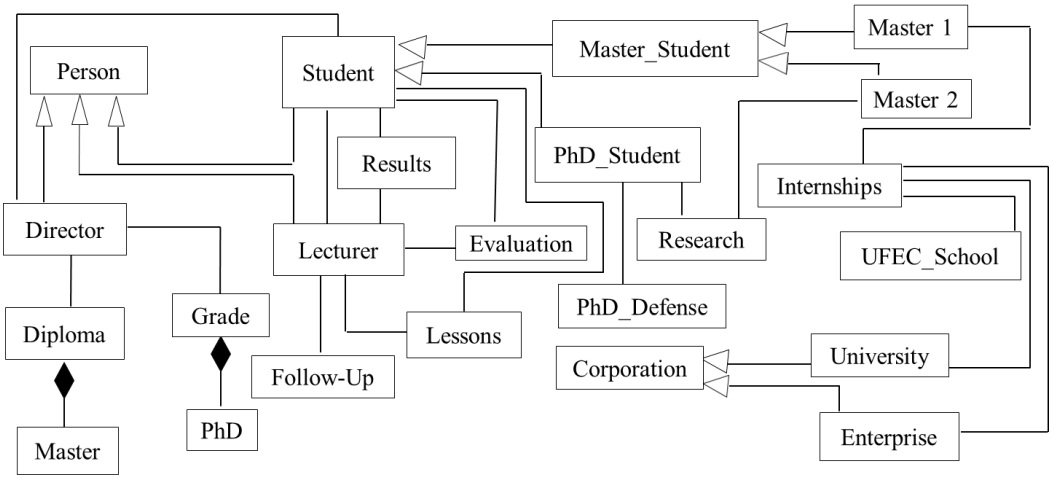


Figure 3: Extract of the 'Class diagram' of the UFEC School "Training Department"

4.3 The dynamical model

The dynamic analysis phase is an important step in the definition of objects and understanding of their functioning. It aims to describe:

- Temporal and event relationships between system objects described in static models.
- Objects state i.e. internal changes during the course of the application depending on the options selected by users.
- Actions performed by objects in a given context.
- External actions of the system on objects in the studied system and the reactions of these objects.

Dynamic modeling is based on several models aimed firstly, to describe the interactions between objects in the system and external systems, and secondly, to study the evolution of internal objects.

We illustrate one type of UML2 diagrams (called sequence diagram or scenario) of our application: the "creation of a new student".

Figure 4 illustrates a part of the dynamical diagram.

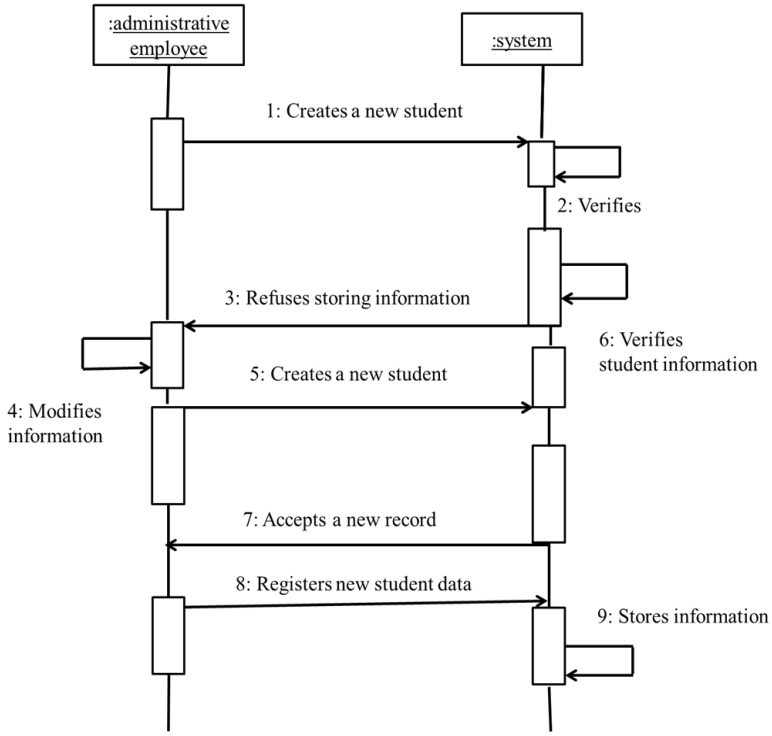


Figure 4: The dynamical diagram ('register a new student')

Section 5 is the conclusion of the different concepts underlined in the paper.

5 Conclusion

In this paper we have developed an Object-Oriented model with the UML formalism in order to implement a real Information System in a graduate school in Paris (France). This work is a part of a type of research conducted in our School. This model has summarized all concepts and knowledge used for the organization of the school. These concepts were formalized with various diagrams according to the Object-Oriented approach with the unified language UML (UML2). The approach permitted us to structure the management of the school which was manual except some electronic files with Excel and Word. The school is a new organization and opened since only four years (2008). In a few years, the school has developed a lot of academic sectors of learning and training. It needs a re-organization and restructuration of knowledge for the management of the school with a better management based on high and modern technologies. The Object-Oriented approach was suitable to manage the complex and various sources of data providing through a multiple of departments. The school has a multimedia department which manages image data. The approach is completely suited. Classes will help us to generate automatically the code and implementation will be easy to complete. We project in a near future to implement the system and place it in the graduate school as soon as possible.

6 References

[1] Mary Culnan. "Mapping the Intellectual Structure of MIS", 1980-1985: A Co-Citation Analysis", MIS Quarterly, Vol. No.11, Issue No.3, (341-353), (Sept 1987).

[2] Peter Keen. "MIS Research: Reference Disciplines and a Cumulative Tradition", in Proceedings of the First International Conference on Information Systems, McLean, E.R. (Ed.), (9-18), (Dec 1980).

[3] Allen Lee. "Architecture as A Reference Discipline for MIS", in Information Systems Research: Contemporary Approaches and Emergent Traditions, H.-E. Nisen, H. K. Klein, and R. A. Hirschheim (eds.), (573-592), (1991).

[4] John Mingers and Frank Stowell. "Information Systems: An Emerging Discipline", (eds.). McGraw- Hill, (1997).

[5] Mark Silver, Lynne Markus and Cynthia Mathis Beath. "The Information Technology Interaction Model: Foundation for the MBA Core Course", MIS Quarterly, Vol. 19, Issue No.3, (361-390), (Sept 1995).

[6] Robert Galliers, Linne Markus and Sue Newell. "Exploring Information Systems Research Approaches", (Eds), (2006).

[7] Claudio Ciborra. "The Labyrinths of Information: Challenging the Wisdom of Systems", Oxford University Press, (2002).

[8] Alan Hevner, Salvatore March, Jinsoo Park and Sudha Ram. "Design Science, in Information Systems Research", MIS Quarterly, Vol. No.28, Issue No.1, (75-105), (2004).

[9] Salvatore March and Gerald Smith. "Design and natural science", in Information Technology (IT), Decision Support Systems, Vol. No.15, (251-266), (1995).

[10] David Avison and Steve Elliot. "Scoping the Discipline of Information Systems", in Avison, D. E. and Pries-Heje, J. (eds), (2005).

[11] Mokrane Bouzeghoub, Georges Gardarin and Patrick Valduriez. "Les Objets ", Eyrolles, (1998).

[12] Laurent Debrauwer and Naouel Karam. "UML2: Entraînez-vous à la modélisation", Editions ENI, (2006).

[13] Kenneth Barclay and John Savage. "Object-Oriented Design with UML and Java", Elsevier Butterworth-Heinemann, (2004).

Firms and Faculty in Convergence: Public Research Organizations and Private Industry Perspective

Muhammad Fiaz
Yang Naiding
Arbab Arsalan
Northwestern Polytechnical University (NPU)
Xi'an, China
fiaz_42@yahoo.com
naiding@nwpu.edu.cn
arbab_09@hotmail.com

Abstract

Alliances and joint ventures are the trade mark of the prevailing competitive and commercialized global world. Research and development is the hall mark of the growing economies and the flourishing multinationals. The huge budgets and state-of-the-art technology demand for joint ventures and alliances for the execution of research and development. Human and capital resource also triggers the technology giants to shake hands with academia and public research organizations for the realization of their research and development endeavors. The past two decades have witnessed the trends of Industry-University partnerships and developed countries have got fruitful results due to these alliances. A similar pattern is observed in the public research organizations for collaborating with the academia for their specialized projects of national defense.

This research contribution is an extension of the academic alliances concept with research and development industry including the public research organizations too. The paper is drafted to highlight the importance of such cooperation while discussing the success stories from the developed world. The case study of Pakistan is presented and the focal point of discussion is the aerospace and communication industry of Pakistan that is still in its incubation state. The developing countries are marked with low research and development budgets and less qualified human resource professionals and this discrepancy leads to the higher need of technological alliances for academics and research achievements. In developing nations, national defense and sovereignty dictates the public organizations to strive for the survival and it compels them for the joint ventures and partnerships with academia and local industry for the achievement of their goals and visions. A comprehensive analysis of nascent aerospace and burgeoning communication industry of Pakistan is presented by the authors and compared with the contemporary industrial giants of the world in the same sector. Guiding principles are drafted for all other developing nations like Pakistan to enhance their alliances with the knowledgeably rich and veterans of research at the academic institutions to achieve innovation and technological milestones.

Key words

R&D – University-Industry Collaboration – Aerospace Industry – Communication Industry – Knowledge Management

Introduction

Commercialization of technology has forced the organizations towards more technical complexities. Technology and trends are changing every day. Growth to global scales of organizations accompanying commercialization has forced the world to rapid growth of scientific knowledge and technological challenges. To meet these challenges, organizations need to make rapid innovations. Consequently, innovation oriented organizations are need to maintain their R&D capacity and keep track of new and competitive advantages. Organizations have been observed to have collaboration tendencies but problem is heavy budget and limited resources to carry out R&D activities. Only a few big firms were able to do that but still they need more knowledge for rapid growth. As a result, firms started to outsource R&D for innovation. R&D collaboration was experiential to be based on type of innovation paradigms. Literature has disclosed two types of paradigms: open innovation paradigm and closed innovation paradigm. In open innovation paradigm, firms like to exploit the innovation while in the close innovation paradigm firms retain their innovation inside the boundaries. R&D collaboration depends upon the openness of the firms that is related to partner's tendency to share the innovative intellectual property [1]. Universities append open innovation paradigm. In spite of the fact that universities protect their innovative knowledgebase and research by patenting but still it is following open paradigm. They have lots of research data that is generated in the form of final year student thesis, research conferences and final project reports. These documents are available overtly and very easy to access. On the other hand, Research labs equipped with latest research materials and professors' expertise in accorded with the latest ideas of research and innovative knowledge are the reasons for business Mogadishu to enhance collaborations with academia. These inducements changed the trends and organizations and industries started to get advantages from this cheap source of external knowledge. Most of the firms' economy depends upon the intellectual property of organization rather than the big spreading pools of tangible assets. That's why organization now focuses on managing their organizational R&D knowledge.

This paper provides a roadmap for the developing countries as Pakistan, India, Bangladesh, Sri Lanka, etc that how these countries can get the advantages from academia. Especially in the field of Communication and Aerospace where R&D is basic requirement of survival but low budget is a major constraint for R&D. Commercialization in these areas can make fast revolutions and causes to raise the GDP of the country. This can be achieved only if these countries support and use the academia research in these fields. Industries and firms have come in touch with academia in India and Pakistan, but still it's far away from the stage where they have to achieve their goals of factual research and overcome innovation challenges. Universities are already collaborating with Public Research Organizations and these research institutions sharing same core competencies for research in Pakistan. If these institutes and organizations promote R&D collaborations in the Aero and Communication sector and establish R&D ecosystems, further developments can be achieved in these sectors more rapidly.

This article begins by emphasizing the importance of R&D collaboration. Second section elaborates the practices made by developed countries for their economic growth and how academia has performed its role to achieve these goals by boosting aero and communication industry up. Roadmap for developed countries have been talked about in next section and future of COMM and Aero industry in the light of academia-industry collaboration makes the heart of this article. Suggestions and recommendations have been positioned in conclusion of this effort.

R&D Collaborations

Major challenges to the organizations are escalating trends of competition, technical complexities due to rapid progress in new product development, increasing R&D Cost, exhausted R&D efforts and limited resources. Considering these challenges, organizations were observed to turn their directions towards using in-home or external R&Ds. For both type of R&Ds, collaboration is very important. Variety of motives are given in the literature for the apparent growth in innovation and technology alliances, and the most important reason for the firms to get-in into collaborative arrangements is innovation. A rational motive for this belief is lack of some or all of the necessary resources including intellectual assets (R&D knowledge) and the risks (technological spillovers) associated with innovation [2]. A strand of literature explains a positive relationship between research intensity and survival rate of the firms [3]. While making collaborations, there are many risks involved and organizations try to avoid alliances. The most important risk is knowledge leakage. Circumstances in which the most proactive alliances are formed and restricted alliance scope even cannot reduce the risk of leakage. Appropriate level of knowledge sharing is required to achieve the collaborative objectives. Emphases of collaborative ventures are mainly risk sharing, cost optimization and managing complex business relationships. Despite the fact that challenges are there, but it's obvious that clustered firms exhibit a higher R&D intensity than their non-clustered counterparts and it enhances their level of technological competence or, more specifically, helps to exploit new R&D opportunities that may arise from localized knowledge spillovers [4].

Communication and Aeronautical industry is considered to be restricted and under state control in developing countries. Due to very heavy and complex nature projects, public organizations avoid to share such projects, especially with universities where people believe in open innovation epitomes. Considering the fruitful results expected from collaboration with academia, modules of such projects can be shared with institutions and better results for innovation and novelty can be achieved in short time.

Trends towards University - Industry R&D Collaborations in Developed Countries

University-Industry Collaboration has been observed to be practiced in many developed countries. Even states are supporting such activities in both service firms as well as in manufacturing sectors. Different empirical researches have been conducted and the results show that public support promotes and encourages the organizations to establish R&D clusters. Core concept behind this effort is to carry out joint R&D activities. Globalization and commercialization of R&D led the firms towards technological complexity. Collaborative R&D networks are the best ways to increase technological intelligence as it provides a common culture for Research and Development based on trust. Considering the importance of articulated R&D environment and benefits of R&D collaborations, developed nations started to ascertain R&D families. States promoted those SMEs and big firms which had already experienced R&D alliances with academia and tended towards academe knowledge reserves. In USA and Europe, governments try to stimulate collaboration between firms, between firms and public institutions. [5]. Similar views were given by Fritsch and Lukas [6] who observed that group of German innovating firms who were engaged in higher levels of product innovation were more likely to establish R&D collaboration. Industrial support of university R&D has risen in the United States since the 1970s. Fig 1 is a clear picture that how industries support for academia R&D has been increased during this period of time. Industrial support currently accounts for 5% of total academic R&D expenditures in United States of America [7].

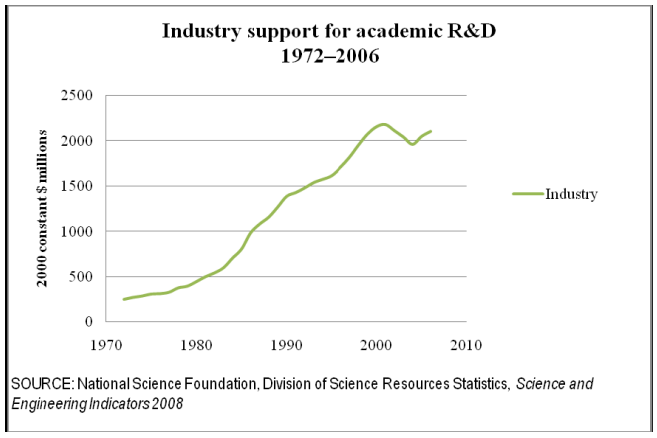


Fig-1: Industry support for academic R&D

University-based research does not automatically evolve into business innovation. A long list of barriers and challenges has been narrated in the literature that hinders the collaboration process. A major challenge among those barriers is the ability of industry to take advantage of university research [8]. Business tycoons and entrepreneurs realized the truth and established gigantic alliances with institutes. A long list of joint projects under an articulated R&D environment has been come up in the directions of science, technology, arts and humanities

USA aeronautical industry is supporting public, private and defense aero plane categories in terms of service and manufacturing [9]. BOEING argues that they are the world's largest manufacturer of commercial jetliners and defense, space and security systems. Even they are considered as top U.S. exporter and company supports airlines in 150 countries. They have a major R&D alliance with GEORGIA TECHNOLOGY, a top research oriented institute in Georgia. Table 1 and 2 sows the list of institutes enjoying the collaborative incentives with industries in USA especially in the field of communication and aeronautics. Government-university-industry research Roundtable (GUIRR) data shows the tendency of different multinational and international firms to collaborate with universities in United States as exhibited in Table-1and 2 [10]

China is the second largest economy in the world with a GDP of totaled $7.8 trillion [11]. Academes are regarded as major contributor for scientific and technological research advancement that leads to creation of innovation and novelty. An outstanding increase in scientific research publications and patenting innovation records are the proofs of Chinese universities R&D efforts and role of industries' support for academe side. Chinese higher educational institutions have more than 25 million students. They have depicted a tremendous increase in research publications and this paper output was increased to 112 000 by 2008.

Industry Partner	University Partner
AGLLENT TECHNOLOGY	UNIVERSITY MICHIGAN
BATELLE	OHIO STATE U
CORNING	PENN. STATE
ELSEVIER	PURDUE
IBM	STANFORD
INTEL	UC BERKELEY
NORTHROP GRUMMAN ELECTRONICS	MIT
SEIMENS	PRINCETON
SEMICONDUCTOR RES CORP.	U. TEXAS AUSTIN
SOUTHWEST RESEARCH INSTITUTE	U TEXAS SAN ANTONIO
HEWLETT-PACKARD (IT)	UCLA

Table-1 (a): Communication Sector

This great achievement distinguished China from other nations as Japan, UK and Germany [12]. China is supporting academia research alliances. Government is also supporting such joint ventures.

Aero industry business jargons have more intentions towards the research at institutes. Harbin Aircraft Manufacturing Corporation, Harbin Embraer Aircraft Industry Ltd and Hongdu Aviation Industry Group are considered as famous brands in aircrafts manufacturers among Chinese aeronautical industry groups.

Industry Partner	University Partner
BOEING	GEORGIA TECHNOLOGY
DEERE	IOWA STATE
LOCKHEED MARTIN	MARYLAND
MARS (Food and Drinks)	UC DAVIS
TEXTRON	NORTHEASTERN
NORTHROP GRUMMAN	CAL TECH
PACIFIC NORTHWEST NAT LAB	WASHINGTON STATE
RAYTHEON	UMASS
NOBLIS	GEORGE WASHINGTON

Table -1 (b) Aero Sector

A long list of communication business mogul is also interested to share projects with universities. Fig-2 shows the increasing level of China in the satellite communication sector. Many Chinese public sector universities are doing projects in this area

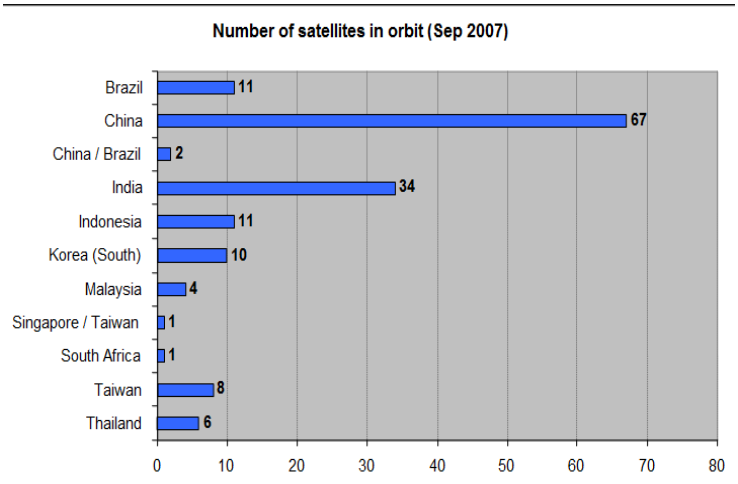


Fig 2: Number of Satellites in orbits (Source: Union of Concerned Scientists Jan. 2008 database)

Trends towards University – Industry R&D Collaborations in Developing Countries

Strategic alliances and collaborative arrangement have been played a prominent role in the field of R&D in all high-tech industries [13; 14]. Developing countries have been observed to spend more expenditure for R&D during last two decades. Ranbaxy is the largest R&D spender in the Indian pharmaceutical industry. It has been reported to spend 4.61% of its sales as R&D expenditure during the year 1994-1995. But it was shot up to 17.21% during 2005-2006. Likewise, second largest R&D spender, Dr Reddys was observed to spend 2.01% of its sales in 1994-1995 but was increased to 10.85% that was due to a fall from 17.12% in the previous year [15]. According to a the data given in Statistical Yearbook for Asia and the Pacific 2011, UNESCO Institute for Statistics, Asian countries have been observed to raise R&D spending tendencies and a significant rise have been observed in few countries as Pakistan, China and Malaysia. Other countries as Singapore, Japan, and Korea are also spending a lot in last decenniums. Gross domestic expenditure on R&D for Asian countries during the period 2001-2008 shows a disappointing condition for the countries as Pakistan and Sri Lanka as shown in the fig-3.

Different factors have been discussed by different authors in the literature that triggered the developed nations towards U-I collaboration. Shorter product life cycle, intense global competition, increasing technological challenges and innovations, unpredictable economic conditions and intensifying the cost of research are major determines that pull the organizations to foster collaboration with academia and research institutions [16;17]. Cutting short of the university budgets, commercialization of academic institutes, need of original research at research labs, hunting jobs and training opportunities, striving research budgets and validation of theoretical research with real life practicalities performed in industry are the major motivators for universities of third-world nations to get the benefits from firms-faculty collaborations.

Case Study: U-I Collaboration for Communication and Aero Industry in Pakistan:

Nuclear concept of this article is to provide a roadmap for developing countries to get benefit from research being conducted at institutes. Author has discussed the case of R&D Collaboration being practiced in an Asian country as Pakistani industries and universities. Mega projects under COMM and Aero category are required more R&D efforts. Strategic organizations avoid sharing their projects due to restricted type of nature of these projects.

The Pakistan Aeronautical Complex (PAC) is aircraft manufacturer in Pakistan. It provides service, assemble and manufacture aircraft for defense. It is the world's seventh largest assembly plant. Similarly, IST, Islamabad is supporting this industry at the academia side. Due to improper planning and ignorance, this industry has not been as flourished as it should be. It needs more intentions for better utilization of researches in this area.

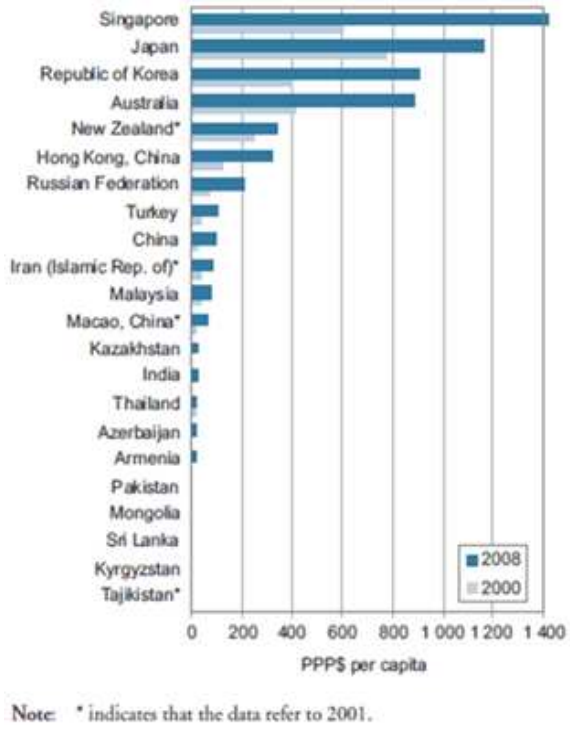


Fig 3: Gross domestic expenditure on Research & Development. Asia and Pacific, PPP$ per

Direct contribution from academia to industries has been greatly appreciated in the advanced economies [18]. Nevertheless, developing countries have lack of wider institutionalized application of such collaborations where industry can get benefit from academia. Reason is less capabilities of innovation tendencies and lack of budgets under research budget heads. Improper establishment of research culture at institutions is also a reason.

University Industry collaboration is a two yaw process. Organizations share their technicalities with the researchers doing research in an absolutely changed environment different from organization culture. In reciprocation, universities share their latest equipped labs, up-to-date and trained researchers' knowledge-base and provide cheap research labor in the form of final year students. In a manner that facilitates these sorts of collaboration are helpful for searching researchers and scientist by giving them opportunity according their interest and field.

Universities and Institutions with special competencies are intended towards establishing collaborations with industries due to the reasons: to enhance research findings, interaction with practitioners, access to technical expertise and practical real world problems. Employment opportunities for university graduates are also become low-priced and convenient [19]. Hurmelinna [20] has added institution reputation as motivating factor among the others. Access to 'empirical data' from industry is also an important factor for collaboration at academe side. To a greater extent, universities have made

liaison cells/R&D cells to hunt, initiate, coordinate and execute joint R&D projects with industries [21].

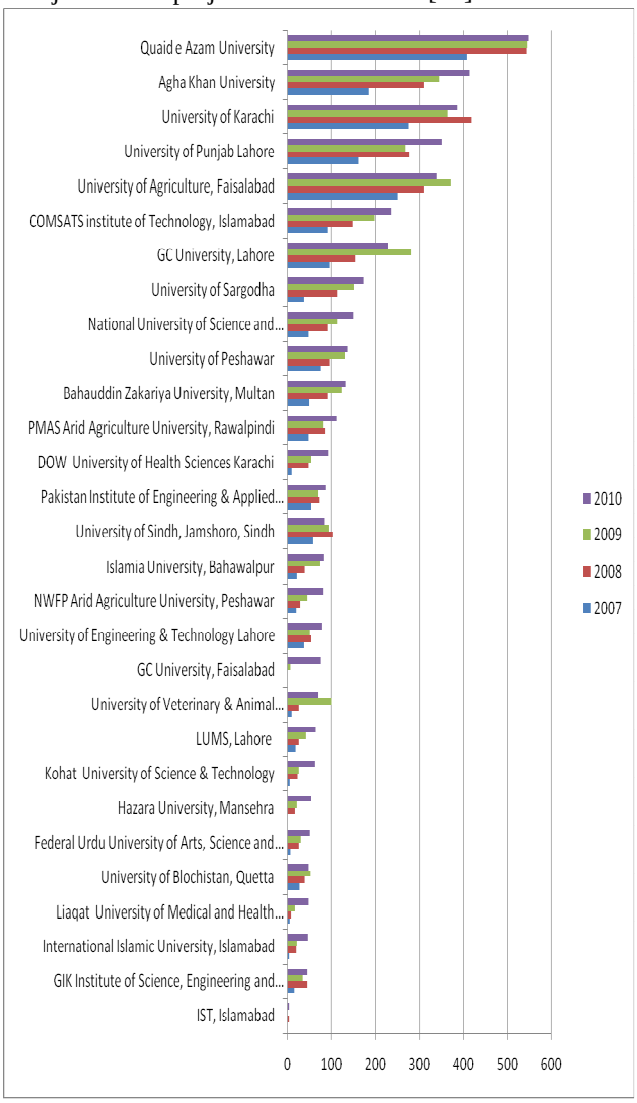


Fig 4: Publications from Pakistani universities appeared in journals indexed by Thomason Reuters as SCI, SSCI and ISI web of knowledge from 2007-2010 (Source: HEC Pakistan website)

There are at last but not the least four research categories of university-industry interactions: research support, cooperative research, knowledge transfer and technology transfer [22]. While making such alliances in the COMM and Aero industry in developing countries as Pakistan, knowledge transfer and technology transfer are major issues to be considered. Reason is different culture and working environment of these two partners.

Literature shows that institutions have played a key role for technological innovations and helped industry to overcome innovative complexities. It has also contributed in the national economy in developed countries. It has already been mentioned that R&D spending has been increased in Pakistan. U-I collaboration journey has been started during past few years. Fig-4 shows the data during the year 2007-2010 when universities in Pakistan have raised the number of research publications in international journals. This increase is miner and requires more emphases in this direction. State support can encourage both partners to enhance this effort. Universities of third-world countries like Pakistan have limited financial resources. In the way as indicated, Institutes with higher studies in Pakistan are seeking more industrial supports in the form of research funds. Quaid e Azam University, Agha Khan University, University of Karachi, University of Punjab Lahore and UET Lahore have joint projects with Communication industry. Higher Education Commission has started various programs and allocated budgets for universities who promotes U-I collaboration. Boom in IT and telecom sector in last decade, has forced the both partners to achieve competitive advantages by making joint R&D efforts. A great rise in the graph of research publications in figure-2 is the indication of increasing research collaborations [23]. Resemblances in the results have also been observed on the webpage of UNESCO Institute for Statistics.

Aeronautical industry invents, assembles and maintains air vehicles. In Pakistan, aerospace field magnitude embraces industrial, commercial enterprises and defense applications. Companies given in list-1 are directly or indirectly participating in Pakistani aerospace industry.

Universities like, IST Islamabad, Pakistan Institute of Engineering & Applied Sciences, Islamabad (PIEAS), University of Punjab Lahore and National University of Science and Technology, Islamabad are collaborating with public sector research institutes in the same field.

Sr. No	Aero Industry
1	NESCOM (National Engineering Scientific Commission)
2	PMO (Project Management Organization)
3	NDC (National Development Complex)
4	SUPARCO (Space and Upper Atmospheric Research Complex)
5	Pakistan Ordinance Factories Wah Cantt.
6	DESTO (Defense Establishment Organization)
7	KRL (Dr. A. Q. Khan Research Laboratories)
8	AWC (Air Weapons Complex).
9	Pakistan Aeronautical Complex (PAF)
10	Scaled Aviation Private Industries Ltd.
11	Integrated Dynamics
12	SATUMA UAV systems

List 1: Aero Industry in Pakistan [24]

List two is about the companies involved in projects of communication sector. Before the economical gravitational pull in Pakistan, IT and Telecom sector was a sky touching industry in Pakistan. A bulk of foreign investment flooded in this land. Universities also stream lined their academicals towards this field.

Sr. no	Telecom Industry
1	Techlogix
2	CyberNet
3	Mindstorm Studios
4	Mobilink Netsol Technologies
5	Ovex Technologies
6	Pakistan Telecommunication Company
7	Palmchip Corporation
8	Southern Networks
9	Telenor Pakistan
10	Nextech Transworld
11	Ufone
12	Warid Telecom
13	Wateen Telecom
14	World CALL

List 2: Communication Industry in Pakistan [25]

Universities like Quaid e Azam University, University of Karachi, COMSATS institute of Technology, Islamabad, University of Engineering & Technology Lahore, LUMS, Lahore and Kohat University of Science & Technology are major academe brands in Pakistan who were engaged in collaborating efforts with most of these industries

Conclusion:

Authors have made an effort to motivate the developing countries to gain the incentives of academia research through this article. By paying very less, developed countries have gained a lot by using university researchers and professors in accord with the hottest ideas. Developing countries like Pakistan, universities of different sectors have made collaborations with some Public Research Organizations and Small & Medium Enterprises, but still many other universities lack this type of alliances and they may follow the proposed R&D Ecosystem which may lead to innovation. This article is also an effort to clear a roadmap for developing countries to make R&D clusters especially in the field of communication and aero industry where in-house R&S is not sufficient to meet the future requirements. If developing countries also follow the same practice and get help from university R&D, more rapid innovations can be made with fewer financial plans. This article is also an insight into communication and aero industry in Pakistan and how different universities have made collaborations with industry but at very minute level. This paper highlights the main aspects that how academia can play a major role for making R&D clusters and how developed countries have got advantages form University R&Ds. In developing countries as Pakistan, still it is a need of time to make such cooperation and relationships among academia and industry for better R&D project execution, especially in the field of aero craft industry and communication sector.

Acknowledgement:

Author is thankful to Northwestern Polytechnical University and China Scholarship Council for supporting this research

References:

[1]. Julio Alberto, 2009, Best practices for University-Industry Research Collaborations), thesis submitted to Massachusetts Institute of Technology.

[2]. Theter BS (2002) Who co-operates for innovation, and why. An empirical analysis. Res Policy 31:947–967

[3]. Cosh, A., Hughes A., Wood E., 1996. Innovation in UK SMEs: Causes and Consequences. ESRC, Centre for Business Researches, University of Cambridge, Mimeo.

[4]. Chang-Yang Lee, Do firms in clusters invest in R&D more intensively? Theory and evidence from multi-country data, Research Policy 38, 1159–1171, 2009

[5]. C. Annique Un Æ Ana M. Romero-Martı´nez, A ´. Montoro-Sa´nchez, Determinants of R&D collaboration of service firms, Serv Bus, 3:373–394, 2009

[6]. Fritsch M, Lukas R (2001) Who cooperates on R&D? Res Policy 30:297–312

[7]. National Science Board. Science and Engineering Indicators. Vol. 1. Two volumes vols. Arlington, VA: National Science Foundation, 2008.

[8]. Lambert, R. (2003) Lambert Review of Business-University Collaboration: Final Report. London: Department of Trade and Industry.

[9]. http://en.wikipedia.org/wiki/Aerospace

[10]. Roberto Fontana, Aldo Geunab, Mireille Matt (2006), Factors affecting university–industry R&D projects: The importance of searching, screening and signaling, Research Policy Vol. 35(2006), Page 309–323

[11]. Economy Watch (2010)

[12]. Adams, J. King, C. Ma, N., (2009), "Global Research Report, China Research and Collaboration in the new geography of science‖, Thomson Reuters

[13]. Hagedoorn, J. and H. van Kranenburg (2003) 'Growth patterns in R&D partnerships: an exploratory statistical study', International Journal of Industrial Organization, 21(4), 517-531.

[14]. Calvert, J. and P. Patel (2003) 'University-industry research collaborations in the UK: bibliometric trends', Science and Public Policy, 30(2), 85-96.

[15]. Sudip Chaudhuri, Is Product Patent Protection Necessary in Developing Countries for Innovation? R&D by Indian Pharmaceutical Companies after TRIPS, WPS No. 614/ September 2007

[16]. Santoro, M., and Chakrabarti, A., (1999). Building industry-university research centers: Some strategic

considerations. International Journal of Management Reviews, 1(3), 225–244.

[17]. Syed Hafeez Ahmad, Conceptual framework for developing strategic partnership between university and industry in Pakistan with particular reference to NWFP , 2nd International Conference on Assessing Quality in Higher Education, 2008, Pakistan

[18]. http://en.wikipedia.org/wiki/Pakistan_Aeronautical_Complex

[19]. National Science Foundation (NSF), 1982b. University-industry research relationships: selected studies. Washington, DC: US Government Printing Office.

[20]. Hurmelinna, P., 2004. Motivations and barriers related to university-industry collaboration-appropriability and the principle for publicity. In: Seminar on Innovation, UC Berkeley.

[21]. National Science Board (NSB), 1993. Science and engineering indicators. Washington, DC: US Government Printing Office.

[22]. Santoro, M., 2000. Success breeds success: the linkage between relationship intensity and tangible outcomes in industry–university collaborative ventures The Journal of High Technology Management Research, 11, 255–273.

[23]. www.hec.gov.pk

[24]. http://theknowledgeworld.com/world-of-aerospace/Pakistan-Aerospace-Companies.htm

[25]. http://en.wikipedia.org/wiki/List_of_companies_of_Pakistan

Address and Participant Entity-Resolution in a Large, Cohort Observational Study Utilizing an Open-source Entity Resolution Tool (OYSTER)

B. Liu[1], U. Topaloglu[1*], W. R. Hogan[1]
[1]Division of Biomedical Informatics; University of Arkansas for Medical Sciences

Abstract - *The National Children's Study (NCS) Arkansas Study Center (ASC) uses the open-source software application called Open sYSTem Entity Resolution (OYSTER), developed at the University of Arkansas at Little Rock (available at [1]), to resolve multiple records of a participant's address. The duplicate records arise because addresses are collected from multiple sources that include instruments, the participant's healthcare provider, and other data-collection forms. The ASC conducts study instruments using the open-source LimeSurvey application. Most address information is obtained via the pregnancy screener instrument, but other instruments also require an address if the subject has moved or plans to move. Participants' demographic information, including address, is entered and managed in caBIG Central Clinical Participant Registry (C3PR). To properly submit participant address and instrument information to Vanguard Data Repository (VDR), we must ensure that a participant's addresses recorded in these applications are resolved if duplicated. Furthermore, given that manual entry of address data in both applications is error prone and subject to variability (e.g., entering St. vs. Street), resolving duplicates is not straightforward and simple string matching will frequently fail to detect duplicates. OYSTER is an entity resolution system that supports probabilistic direct matching, transitive linking, and asserted linking. To facilitate prospecting for match candidates (blocking), the system builds and maintains an in-memory index of attribute values to identities. Once OYSTER identifies the duplicates, we manually resolve them in LimeSurvey and C3PR, and we are moving to an automated process.*

Keywords: Participant address data, entity resolution, OYSTER

1 Background

The National Children's Study Arkansas Study Center (ASC) conducts study instruments using the open-source application LimeSurvey. The address information is obtained via the pregnancy screener, and then entered and managed in caBIG Central Clinical Participant Registry (C3PR). Because address data are collected from multiple sources that include instruments, the participant's healthcare provider, and other data-collection forms, duplicate records arise. To properly submit participant address and instrument information to Vanguard Data Repository (VDR), we must ensure that a participant's address records in these applications are resolved if duplicated. Furthermore, given that entry of address data in both applications is error prone and subject to variability (e.g., entering St. vs. Street), resolving duplicates is not straightforward and simple string matching will frequently fail to detect duplicates. Based on this requirement, an Entity Resolution (ER) process is applied on current study.

Entity Resolution is the process of determining whether two references to real-world objects are referring to the same object or to different objects [2]. Real-world objects can be identified by their attributes and by their relationships with other entities. Most Entity Resolution systems take entity references and base their decisions upon the degree to which two references have similar attribute values, i.e. "matching" to make this decision. ER has also been studied under other names including but not limited to record linkage [3], deduplication [4], reference reconciliation [5], and object identification [6].

Entity resolution systems can be broken out into four categories based on their underlying architectures [1]: Merge-Purge, Heterogeneous database join, Identity Resolution, and Identity Capture. In Merge-Purge, entity references are systematically compared to each other and separated into clusters (subsets) of equivalent records. It is the most common form of ER and also known as record linkage. In Heterogeneous database join, a transactional ER system where attribute values from an input reference is translated into queries to different databases and database tables. The query results are analyzed to determine if there are references in databases that are equivalent to the input reference. In Identity Resolution, all incoming references are resolved against a predefined set of managed identities and each identity in an identity resolution system has a fixed identifier that can be used to link references that are equivalent to the identity, thus creating a persistent link. Notice that in Heterogeneous database join, the reference attribute values are compared and in Identity Resolution the reference identifiers are compared. Identity Capture is a type of identity resolution in which the system builds (learns) a

* University Tower 400, 1123 S. University Ave., Little Rock, AR 72204 UTopaloglu@uams.edu

set of identities from the references it processes rather than starting with a known set of identities.

Regardless of the architecture, ER systems typically apply four techniques for determining equivalence: direct matching, transitive equivalence, relationship resolution, and asserted equivalence. Direct matching determines equivalence between two references based on the degree of similarity between the values of corresponding attributes. The five basic methods for determining the similarity of attribute values are: exact match, numerical difference, approximate syntactic match, approximate semantic match, and derived match codes. Transitive equivalence determines equivalence through the use of an intermediary or common record. This means that if record A is equivalent to record B, and record B is equivalent to record C, then through transitive equivalence, record A could be linked to record C. Relationship resolution determines equivalence by exploring patterns of relationships among references that do not rise to the level of equivalence. This is often done through the use of techniques borrowed from graph theory and network analysis. Relationship resolution allows multiple relationships to be considered and multiple equivalence decisions to be made at the same time. Asserted equivalence is the instantiation of a link between two references based on a prior knowledge that they are equivalent. An asserted equivalence often takes the form of one record carrying the attribute values of two non-matching references. This means a single reference may contain two first names and two last names that do not match but based on prior knowledge, such as the person changed his name, both first and last names refer to the same real-world entity.

OYSTER is an open-source software development project sponsored by the Entity Resolution and Information Quality (ERIQ) Research Center at the University of Arkansas at Little Rock [1]. OYSTER (Open sYSTem Entity Resolution) is an entity resolution system that supports probabilistic direct matching, transitive linking, and asserted linking. To facilitate prospecting for match candidates (blocking), the system builds and maintains an in-memory index of attribute values to identities. Because OYSTER has an identity management system, it also supports persistent identity identifiers. OYSTER is written in Java and the source code and documentation are available as a free download from the ERIQ website [1] for use under the OYSTER open-source license. Although the original version of OYSTER was developed to support entity resolution for student records in longitudinal studies, the system design readily accommodates a broad range of ER domains and entity types. A key feature of the system is that all entity and reference-specific information is interpreted at run-time through user-defined XML scripts. This allows OYSTER to be configured as a merge-purge, identity capture, or identity resolution system. OYSTER does not use an internal database for its operation, and system inputs and outputs can either be text files or database tables. XML scripts are used to define (1) all entity identity attributes, (2) the layout of each reference source, and (3) the identity rules for resolving each reference source.

2 Methods

In this paper, we report our results applying OYSTER to the ER problem with participant address data. To protect participants' privacy, all personally-identifying data were blocked or replaced by dummy values.

Since OYSTER does not use internal database our first step is to transfer participant data from C3PR to local computer and input into OYSTER. The database table is shown as figure 1 (actually personally-identifying data has been blocked out for privacy reasons). Three XML files are required to be edited before we start OYSTER operation. The RunScript file (figure 2) specifies the path and name of the Input Descriptor file, the Input Attributes file, the Output Identity file, and the Output Link file. The Input Descriptor file (figure 3) specifies the location of the source data which was transferred from C3PR, and specifies the attribute locations so OYSTER can read different attribute values from different locations. The Input Attributes file (figure 4) defines the matching rules.

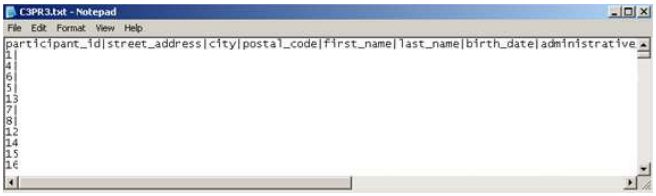


Figure 1 - Participant address data transferred from C3PR to local computer (all personally-identifying data has been blocked out for privacy reasons).

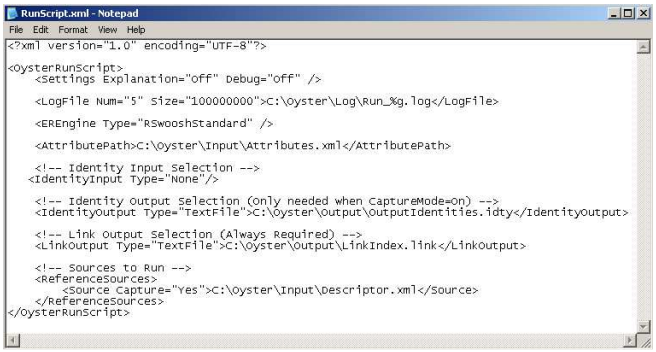


Figure 2 - XML file RunScript.

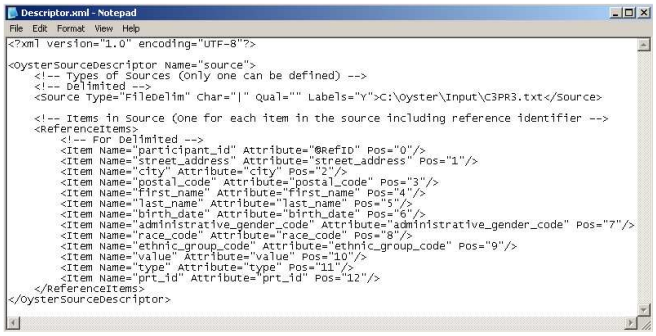


Figure 3 - XML file Descriptor.

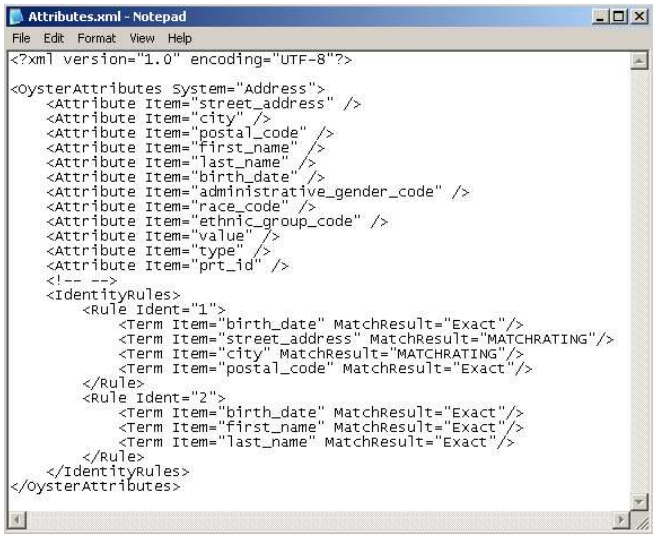


Figure 4 - XML file Attributes.

When OYSTER clusters multiple references, it indicates that according to the matching rules in the Input Attributes file, multiple references likely refer to one real-world entity. We manually investigated all clusters of size greater than one (meaning that there were potential duplicate references). Since the matching rules may not cover all possible duplicates, it is necessary to adjust the matching rules, rerun OYSTER operations, and review again all clusters of size >1. In this paper, we identified four types of duplicates: two true duplicates, cross-column equivalent records and missing value equivalent records; two false duplicates, attribute occupancy value false equivalent records and shared attribute value false equivalent records. The cross-column equivalent records describe equivalent attribute values exist cross two or more attributes and all rest attribute values are exactly matching between two references. For example, reference A and B have six attributes in which the third attribute of A is matching the fourth attribute of B and the fourth attribute of A is matching the third attribute of B, and the rest attributes (first, second, fifth and sixth) are matching each others between A and B. The missing value equivalent records describe two references have equivalent values on all attributes except one reference missed one or more attribute values. The attribute occupancy value false equivalent records describe that multiple references have one or more equivalent attribute values occupied by dummy values and non-matching values on all rest attributes. The shared attribute value false equivalent records describe two references shared same values on one or more attributes but have non-matching values on all rest attributes.

3 Results and Discussion

The output of OYSTER after processing the NCS address data are shown in figure 5. OYSTER processed 289 references, of which 288 were valid. One reference was created during the Extract Transform Load (ETL) process. OYSTER clustered the 288 remaining references into 281 clusters. Of the 281 clusters, 277 had one reference (cluster size of one), indicating that they are unique. Four clusters have contained more than one reference (i.e. 2 clusters with 2 references, 1 cluster with 3 references, and 1 cluster with 4 references), indicating that duplicates might exist within these clusters. The output LinkIndex file (figure 6) also confirmed the results. The three columns in the LinkIndex file are RefID, OysterID and rule. The RefID field specifies the location in the input file where OYSTER found the reference. The OysterID is the unique identifier OYSTER applied to each cluster, and can be used to review duplicates. The rule value indicates which matching rule was applied on current operation. A null value indicates unique reference, and a non-null value signifies possible duplicates. The remainder of the results section describes our analysis of these potential duplicates.

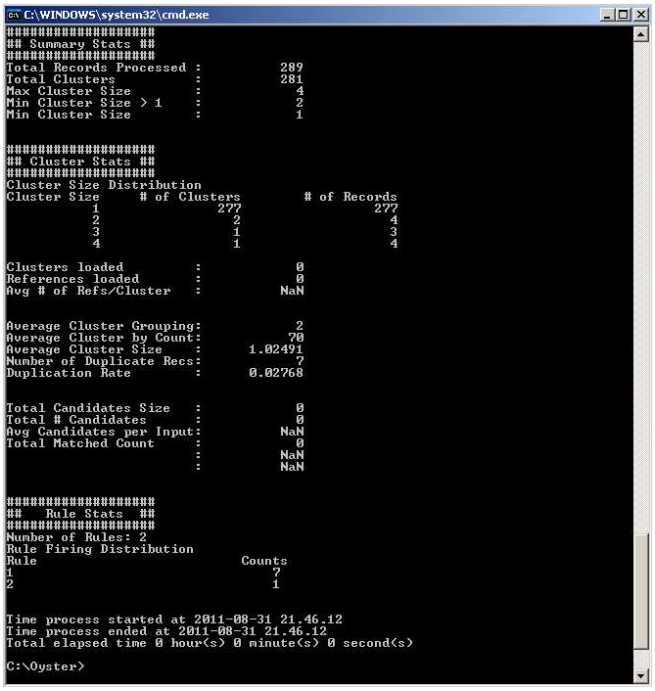


Figure 5 - OYSTER operation result.

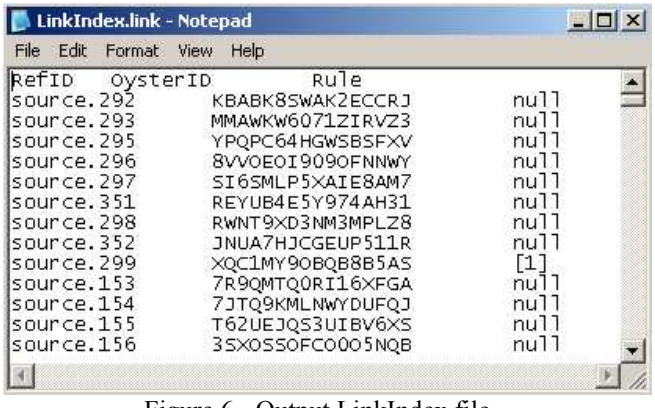


Figure 6 - Output LinkIndex file.

3.1 Cross-column equivalent records.

The first type of potential duplicate was cross-column equivalent records (Figure 7). Further study allowed us to identify that one size two cluster belongs to this type in which one participant's information was entered into the system twice and the two records switched first name and last name.

	participant_id	street_address	city	postal_code	first_name	last_name	birth_date
9	21	123 N Oak St	Little Rock	72200	Mary	Smith	02/17/1982 0:00
10	46	123 N. OAK STREET	LITTLE ROCK	72200	Smith	Mary	02/17/1982 0:00

Figure 7 - Cross-column equivalent records (all personally-identifying data were replaced by dummy values).

3.2 Missing value equivalent records.

The second type of potential duplicate that we studied was missing value equivalent records (Figure 8). One size two cluster belongs to this type in which one participant's information was entered into system twice, but for one reference, all attributes have values and for the other reference, three attributes had a missing value.

	participant_id	street_address	city	postal_code	first_name	last_name	birth_date
156	362	123 N Oak St	Little Rock	72200	Mary	Smith	02/17/1982 0:00
157	362				Mary	Smith	02/17/1982 0:00

Figure 8 - Missing value equivalent records (all personally-identifying data were replaced by dummy values).

3.3 Attribute occupancy value false equivalent records.

The third type of potential duplicate that we studied was attribute occupancy value false equivalent records (Figure 9). We discovered that one size five cluster belongs to this type in which five participants shared the same date of birth, but the value of this attribute for all five records was nonsensical (01/01/1000). We presume that the data collectors must have missed all five participant's date of birth and had to replace the missing value by an occupancy value to upload the participant's information to the system successfully.

	participant_id	street_address	city	postal_code	first_name	last_name	birth_date
133	172	123 N Oak St	LittleRock Rd	72200	Mary	Smith	01/01/1000 0:00
134	180	456 Main St	Little Rock	72200	John	Don	01/01/1000 0:00
135	243	466 Main St.	Little Rock	72200	Jane	Don	01/01/1000 0:00
136	363	100 Broad Rd	Little Rock	72200	David	Smith	01/01/1000 0:00
137	364	666 Main St	Little Rock	72200	Mary	Don	01/01/1000 0:00

Figure 9 – Attribute occupancy value false equivalent records (all personally-identifying data were replaced by dummy values).

3.4 Shared attribute value false equivalent records.

The fourth type of potential duplicate we studied was shared attribute value false equivalent records (Figure 10). One size two cluster belongs to this type in which two participants shared same date of birth and two records are not duplicates.

	participant_id	street_address	city	postal_code	first_name	last_name	birth_date
228	299	123 N Oak St	Little Rock	72200	Mary	Smith	02/17/1982 0:00
229	300	456 Main St	Little Rock	72200	John	Don	02/17/1982 0:00

Figure 10 - Shared attribute value false equivalent records (all personally-identifying data were replaced by dummy values).

3.5 Summary of analysis of potential duplicates.

Of the four clusters, two represented true duplicates (two references referred to same entities in the real world). The other two clusters included false matches: one because of errant data and one because of a true, shared attribute.

4 Conclusion

We applied the OYSTER open-source entity resolution tool to participant address data. We detected two true duplicates, with two false positives.

This work demonstrates the value of entity resolution tools even when the size of datasets is relatively small. Although we had less than 300 study participants at the time of our analysis, we detected two duplicates. The tool also detected errant data, albeit through false matches that had to be reviewed manually.

Future work includes automating the process so that cluster detection is done automatically on a periodic basis. Also, the OYSTER tool does not currently have cross-column detection functionality, and we plan to add that functionality.

5 Acknowledgement

This work has been supported partly by the award #HHSN275200800026C from Dept. of Health and Human Services and by the award #1UL1RR029884 from the National Center for Research Resources

6 References

[1] http://ualr.edu/eriq/downloads/

[2] John Talburt. "Entity Resolution and Information Quality". Morgan Kaufmann, 2011.

[3] HB Newcombe, JM Kennedy, SJ Axford, and AP James. "Automatic linkage of vital records"; Science, 130:954–9, 1959.

[4] S. Sarawagi and A. Bhamidipaty. "Interactive deduplication using active learning"; In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 269–278. ACM New York, NY, USA, 2002.

[5] X. Dong, A. Halevy, and J. Madhavan. "Reference reconciliation in complex information spaces"; In Proceedings of the 2005 ACM SIGMOD international conference on Management of data, pages 85–96. ACM New York, NY, USA, 2005.

[6] Perrochon, Louis. "Translation Servers: Gateways Between Stateless and Stateful Information Systems"; In Proceedings of the Network Service Conference 1994.