Data Profiling Using Attribute Clustering

M. Heidi McClure
The University of Sheffield, and
Intelligent Software Solutions, Inc
5450 Tech Center Dr., Suite 400
Colorado Springs, CO 80919

Abstract— *Finding trends in database data is hard when presented with data sets containing many attributes (columns). The difficulty is increased when the data is in text fields and may include large summary or remarks fields. This paper discusses an approach that uses attribute level clustering in order to discover trends or profiles in the data. This is different from traditional uses of clustering in that each attribute is clustered separately and then the results are combined to define profiles. For example, in a case study of the Global Terrorism Database (GTD) data set, there are 98 columns (attributes) in the data. A profile might be defined by a particular group, attack type, weapon type and by specific information found in larger remarks-type fields. The profiles will show the values of these attributes along with all the records that matched that profile.*

Keywords: attribute-clustering, WebTAS, clustering, GTD, visualization, data-profiling

1. Introduction[1]

A requirement from a customer is to discover profiles of related objects in a database. The objects are from a table that may have many attributes and may have one-to-one or one-to-many joins to other tables. Data from attributes[2] from the top level table and all joined in tables may be considered for profiling. As noted in the abstract, finding trends in database data is hard when the data contains many columns and when that data includes large text fields like summaries, notes or remarks fields.[1] The solution presented in this paper brings together clustering algorithms and link analysis displays to discover profiles in the data.

The system on which to build this profiling discovery capability is the Web-enabled Temporal Analysis System (WebTAS) - a US government off-the-shelf tool used for data integration, visualization and analysis [2]. Attribute clustering allows for the discovery of profiles. Profiles help users (customers) make sense of their data.

This paper discusses the details of an attribute clustering implementation built on the WebTAS platform and it discusses areas for further research. It also presents a visualization using clustering in a link analysis chart display that includes visual clustering algorithms.

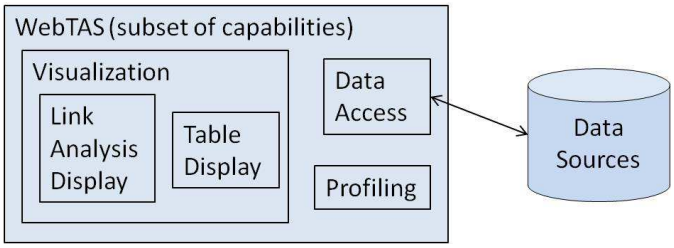

Fig. 1: General Architecture

2. Background

This section will describe the building blocks of the profiling system.

2.1 WebTAS

WebTAS accesses data from any traditional relational database like SQLServer, Oracle, etc and WebTAS access many other sources of data - such as file system data, live streams of data and web services. When accessing these other sources, a custom data source capability found in WebTAS is used. WebTAS's strength lies in its ability to visualize data from disparate sources on tables, graphs, timelines, grids and link analysis charts. For profiling, a custom data source has been added to WebTAS that allows a specific query to be performed that produces a result set and then attribute clustering to be performed on that set. Once profile results are available, link analysis is used to display the results. Link Analysis contains some visual clustering algorithms which further group like profiles together providing another level of clustering of the results.[2], [3]

2.2 Attribute Clustering (Profiling)

Data mining (or text mining) clustering algorithms are usually applied to documents, bodies of text or other collections of text and provide results that group or cluster documents or records into buckets. Clustering of this type applies one

[1]Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

[2]In this paper, attributes, fields and columns are used interchangeably

Table 1: Clustering Algorithms Available

Algorithm	Description
lingo	Works well with large text fields. Records may be a member of multiple clusters. Has descriptive names for clusters
distinct	Like SQL distinct - full field matching - records may be only in one cluster
katz spatial	Geospatial clustering algorithm. Only works on location attributes
katz	each record assigned to single cluster. Uses linear programming to determine cluster centroids
lda	Latent Dirichlet Allocation - each record assigned to a single cluster

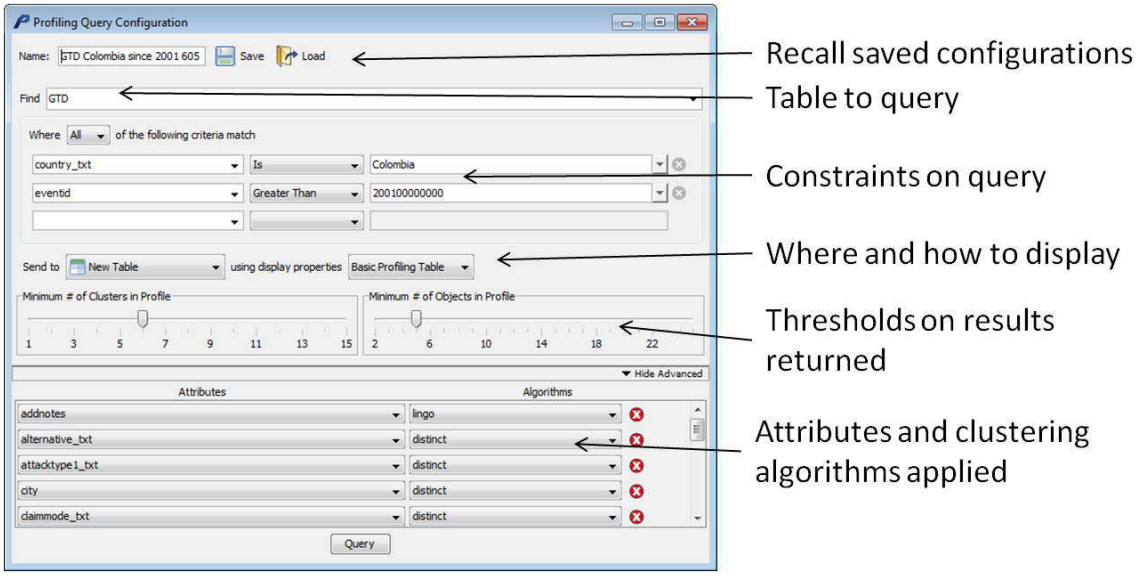

Fig. 2: Profiling Configuration UI

algorithm per pass to the set of data [4], [5]. Attribute clustering applies clustering algorithms to columns or attributes of data, usually pulled from databases. The clustering algorithms applied may be customized to best suit the type of data being clustered. To form profiles, records are grouped based on membership in the same attribute level clusters.

Attribute clustering seeks to discover profiles[3]. Clustering groups similar objects or records into groups or buckets. While an exact definition of clustering is not available, a range of clustering techniques may be found in Xu and Wuncsh's survey of clustering algorithms[5]. As described there, clustering may place records into one and only one group or clustering may place records into multiple buckets.

When data consists of many attributes (that is, many features or columns), one clustering algorithm may be more appropriate than another for a specific attribute. For example, when data consists of short pick-list driven data, a simple grouping algorithm like an SQL distinct function or group by call [6] may be all that is needed. If an attribute is a geographical coordinate, a geo-clustering algorithm is appropriate which will group objects based on how close they are physically to each other. For large text fields, an algorithm like lingo [7] that allows a record to be a member in one or more clusters is appropriate. Numerical data requires yet another clustering algorithm [5].

Attribute clustering is a way to cluster each attribute separately using specialized clustering algorithms and then to bring the results together based on membership in same attribute clusters. In the approach presented in this paper, a minimum number of matching clusters is selected along with a minimum number of matching objects in the profile.

2.2.1 Clustering Algorithms

The clustering algorithms used in this case study are briefly described in Table 1. For more information on their specifics, please see [8], [5], [4].

[3]the verb profiling is used interchangeably with attribute clustering in this paper

2.2.2 Profiling Configuration Dialog

Using the WebTAS infrastructure, a custom data source is used for profiling. This customer data source uses a very large and complicated query string - similar to a large SQL statement, but customized for use with WebTAS and with profiling. In order to hide the ugliness of the large query, a user interface has been created to allow easier selection of profiling criteria. See Figure 2.

2.2.3 Example of Attribute Clustering

As a simple example, consider seven records that are sent to attribute clustering. Attribute one (a1) generates three clusters (a, b, c); Attribute two (a2) generates two clusters (d, e); Attribute three (a3) generates four clusters (f, g, h, i). The gray boxes around the results is to highlight the records in those clusters. '*' character indicates the record fell into the specified cluster for the specified attribute. See Table 2.

Table 2: Clusters Found by Attribute

attributes->		a1			a2		a3			
	cluster ->	a	b	c	d	e	f	g	h	i
records ->	r1	*	*		*		*	*		
	r2	*	*		*		*			
	r3		*	*	*	*			*	*
	r4			*					*	*
	r5					*		*		*
	r6	*	*		*			*		
	r7		*		*					

Examining pairs of records, we find the following and could conclude that r1-r2 and r1-r6 pairs are strongly related. See Table 3.

By looking at the records that match each set of matching clusters, we find that clusters b and d together are found to match five records. This, in addition to the "a,b,d,f" and the "a,b,d,g" cluster sets, form the highest strength clusters shown highlighted. See Table 4.

2.3 Visualization

Link analysis of the profiles and their member objects presents a good visualization of how objects are related. WebTAS link analysis has four visual clustering algorithms[2]

- Springs and Repulsion (nodes repel, links attract)
- Clustering - Self Organizing (Fruchterman-Rheingold Algorithm)
- Filling ISOM (Inverted Self Organizing Map)
- Balanced (Kamada-Kawai Algorithm)

The Clustering - Self Organizing visual clustering algorithm works best for visualizing profiles. See Figures 3 and 4.

Details of the profiling results may also be displayed to a table. See Figure 5.

Table 3: Record Pairs

record pair		matching clusters	number clusters match	
r1	r2	a, b, d, f	4	strongly related
r1	r3	b, d	2	weaker since only two clusters match
r1	r4		0	not related
r1	r5	g	1	weak
r1	r6	a, b, d, g	4	strongly related
r1	r7	b, d	2	weaker since only two clusters match
r2	r3	b, d	2	weaker since only two clusters match
r2	r4		0	not related
r2	r5		0	not related
r2	r6	a, b, d	3	related
r2	r7	b, d	2	weaker since only two clusters match
r3	r4	c, h, i	3	related
r3	r5	e, i	2	weaker since only two clusters match
r3	r6	b,d	2	weaker since only two clusters match
r3	r7	b,d	2	weaker since only two clusters match
r4	r5	i	1	weak
r4	r6		0	not related
r4	r7		0	not related
r5	r6	g	1	weak
r5	r7		0	not related
r6	r7	b,d	2	weaker since only two clusters match

Table 4: Discovered Data Profiles

profile candidates	records match	total records	
a, b, d, f	r1, r2	2	higher strength because more clusters in profile
a, b, d, g	r1, r6	2	higher strength because more clusters in profile
b, d	r1, r2, r3, r6, r7	5	higher strength because more records match profile
g	r1, r5, r6	3	
a, b, d	r2, r6	2	
c, h, i	r3, r4	2	
e, i	r3, r5	2	
i	r4, r5	2	

3. Results

Tests have been performed on the Global Terrorism Database (GTD)[9]. The case study presented here is for the country of Colombia from the year 2001 thru 2008 (includes approximately 600 records). Seventeen (17) attributes are used for profiling the data - the attributes were chosen if they contained data in the records selected. The limits of a minimum of 6 clusters and 6 records are used since they presented a manageable number of profile results (approx 150) for the link analysis display. The goal is to see if data profiling can discover knowledge in the data not easily found in other ways.

A sample of the data sent to a table may be seen in Figure 5. The Profile Summary field shows attribute names and their values for all attributes that matched. The second column shows all the records or data objects that matched the data profile discovered. They are links to the details of the records. The summary information for the GTD records include city, country, event id and province or state.

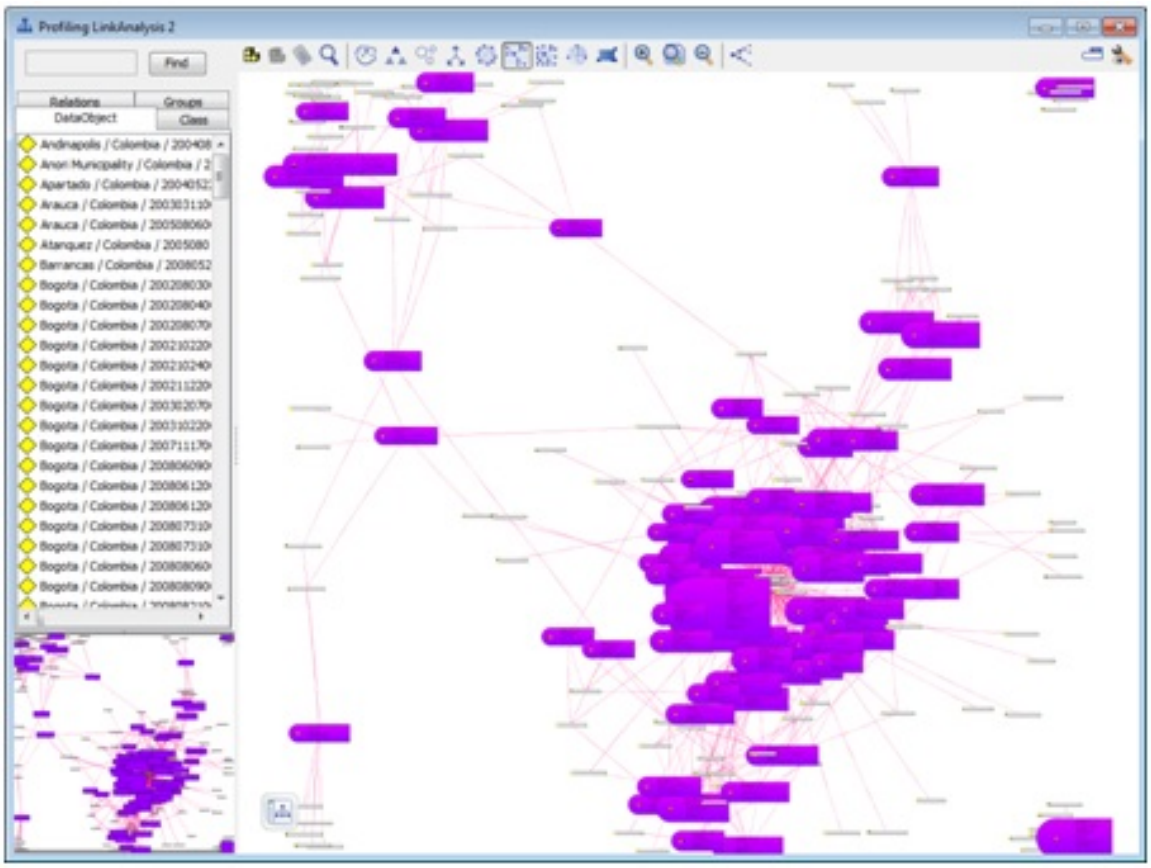

Fig. 3: Clustered Link Analysis Display

Initial display to link analysis may be seen in Figure 6. (The purple or dark icons are the profiles, the light gray are the member records.) By default, data is displayed to the link analysis charts using a radial tree layout. Link analysis shows objects and their related objects - for example, sending profile results to a link chart will show the profile objects (purple or dark) in the first ring of objects and linked to each profile object in the 2nd radial row are the report objects (gray) that are in that profile. Using the radial tree layout, you see that there are some report objects that are members of many profile objects.

Performing Fruchterman-Rheingold Algorithm for visual, self organizing clustering is shown in Figure 3. The Fruchterman-Rheingold algorithm is an example of a force directed layout algorithm where nodes repel and attract - the result is that the profiles and their member objects cluster visually based on how related the profiles are and which objects are members of the profiles.

Notice the separated purple (dark) nodes - these are profiles that don't share many objects with the nodes in the large grouping in the center of the chart. They describe profiles which have distinct characteristics which are not shared by other profiles in the link chart display.

The details of the profile in the upper left of the link chart may be seen in Figure 4. The profile has grouped records for bombing/explosion attack types by the National Liberation Army of Colombia (ELN) that reported injuries and was targeting a business.

Table 5 shows descriptions of some of the other profiles discovered using attribute clustering.

4. Conclusion

The discovered profiles enhance understanding of the data but also allow the customer to categorize new data and more quickly know if the new data matches a pattern which has been seen before. They may also use data in the discovered profiles to know how events are related. Discovered profiles may be examined to know how to prevent the same events from happening again.

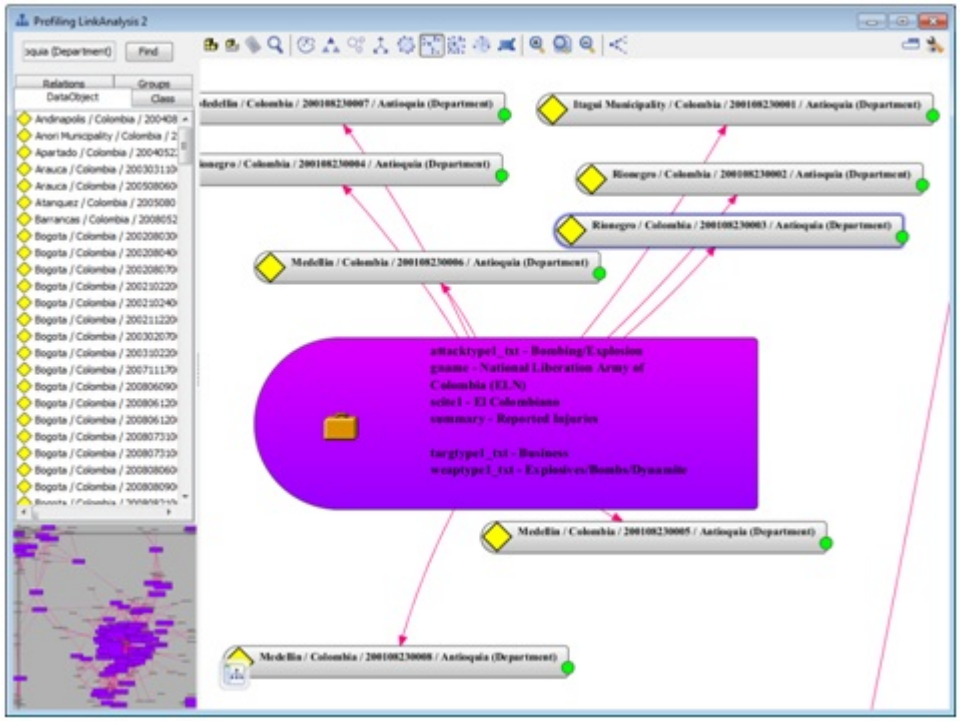

Fig. 4: Zoomed In Link Analysis

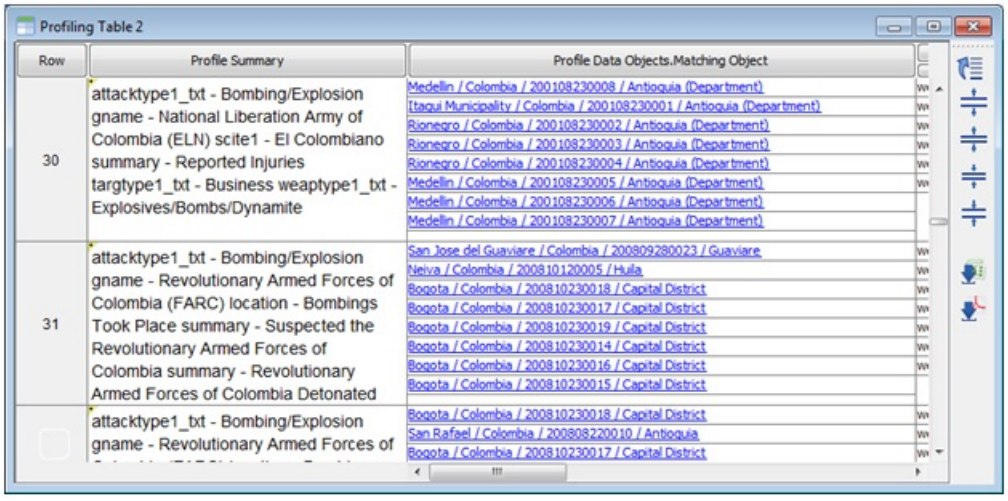

Fig. 5: Table Display of Profiles

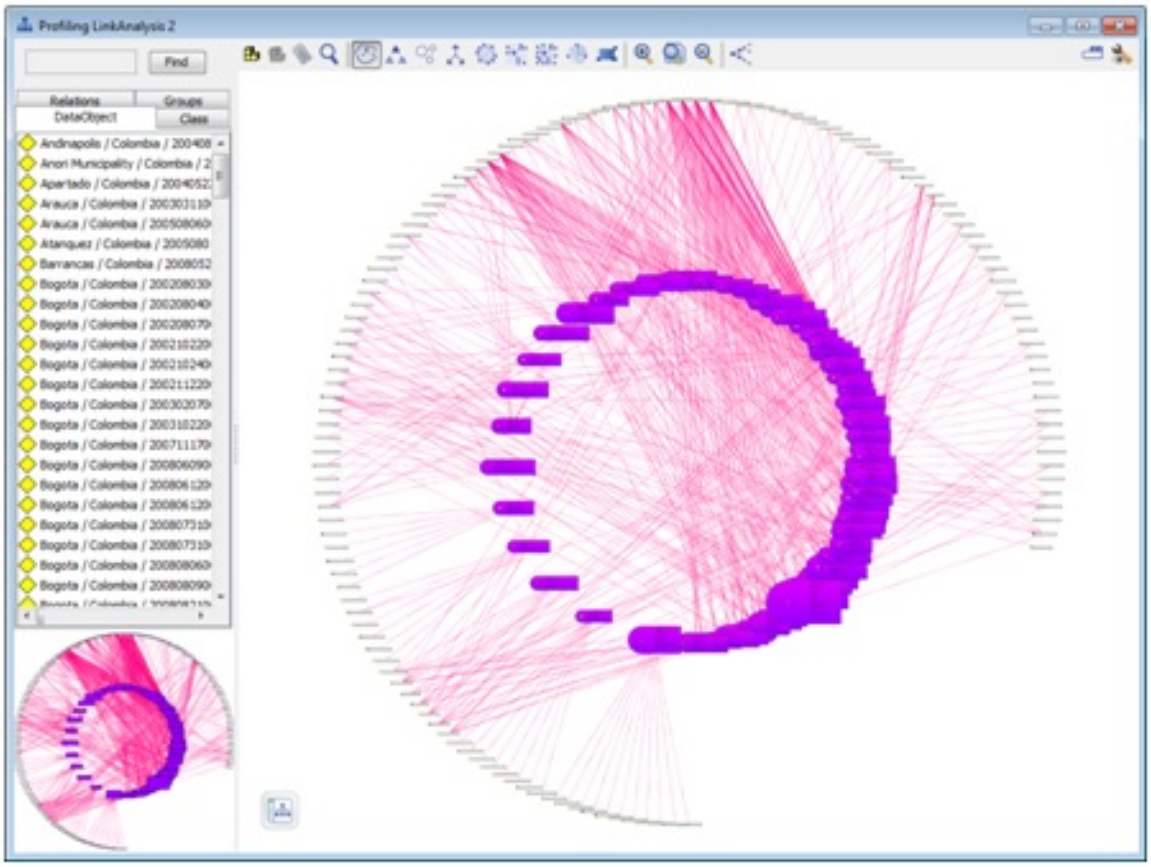

Fig. 6: Link Analysis Circular Display

Table 5: Details of Link Analysis Chart

Where in link chart	Details of profile
Upper left	attacktype1_txt - Bombing/Explosion gname - National Liberation Army of Colombia (ELN) scite1 - El Colombiano summary - Reported Injuries targtype1_txt - Business weaptype1_txt - Explosives/Bombs/Dynamite
Upper right	attacktype1_txt - Armed Assault gname - Revolutionary Armed Forces of Colombia (FARC) summary - Killed by the Revolutionary summary - Suspected the Revolutionary Armed Forces of Colombia summary - Members of the Revolutionary Armed Forces weaptype1_txt - Firearms
Lower left	addnotes - Marked the 40th Anniversary of its Founding addnotes - Founding addnotes - Police attacktype1_txt - Bombing/Explosion gname - Revolutionary Armed Forces of Colombia (FARC) weaptype1_txt - Explosives/Bombs/Dynamite
Lower right	attacktype1_txt - Bombing/Explosion city - Puerto Colon corp1 - Colombian Petroleum Enterprise (Ecopetrol) gname - Revolutionary Armed Forces of Colombia (FARC) location - Bombings Took Place location - Villages of Puerto Colon and San Miguel scite1 - El Tiempo scite1 - January 2 summary - Villages of Puerto Colon and San Miguel summary - Fuerzas Armadas Revolucionarias de Colombia FARC Guerrillas targtype1_txt - Utilities weaptype1_txt - Explosives/Bombs/Dynamite

Although the results shown in this paper are from the GTD, there is nothing preventing profiling from being run on any data that WebTAS can see. As an additional experiment, profiling was run on a table where similar or even duplicate records exist. Profiling was able to identify and link these related records.

5. Future

5.1 Classification (Categorization)

Once interesting profiles are discovered, these related records may be used to train a classifier (categorization in WebTAS). Then when records are inspected by the system, they may be classified into buckets based on the profiles trained. [4]

5.2 Entity Extraction

Entity extraction in this context is the process of pulling entity information out of text. Term extraction may be a better way to think of this kind of entity extraction [8]. An enhancement is to perform entity extraction on data and then apply attribute clustering on the extracted entities. (this has been prototyped, but refinement to the entity extraction grammars has yet to be done.)

5.3 SEER

Once characteristics of a profile are discovered, these characteristics may be incorporated into a WebTAS SEER model for detection of new records matching defined profiles. Situation Exploitation Engine Real-time (SEER) is a component of WebTAS which allows detection of patterns in data both on historical and in a near real-time manner.[3]

Acknowledgments

Work described in this paper was funded at various times by Air Force Research Laboratory (AFRL) at Rome, NY (AFRL Contract FA8750-06-D-0005-0012) and the Defense Advanced Research Projects Agency (DARPA). Some of the work came from the Personalized Assistant that Learns (PAL) program, a DARPA program that is currently in military transition.

The author thanks her advisor, Dr. Mark Stevenson, of The University of Sheffield for being a sounding board during the development of this solution. The author also thanks her brother, Dr. John McClure, for his helpful reviews of this paper and Intelligent Software Solutions, her employer, for presenting such an interesting problem to solve.

The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

[1] S. Džeroski, "Multi-relational data mining: an introduction," *SIGKDD Explor. Newsl.*, vol. 5, pp. 1–16, July 2003. [Online]. Available: http://doi.acm.org/10.1145/959242.959245

[2] (2011) Webtas overview. Intelligent Software Solutions. Intelligent Software Solutions - http://www.issinc.com/solutions/webtas-overview.html. [Online]. Available: http://www.issinc.com/solutions/webtas-overview.html

[3] M. Gerken, R. Pavlik, C. Houghton, K. Daly, and L. Jesse, "Situation awareness using heterogeneous models," in *Collaborative Technologies and Systems (CTS), 2010 International Symposium on*, may 2010, pp. 563 – 572.

[4] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Burlington, MA: Morgan Kaufmann, 2011.

[5] R. Xu and D. W. II, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, vol. 16, pp. 645–678, 2005.

[6] K. Kline, D. Kline, and B. Hunt, *SQL in a Nutshell*, 3rd ed. O'Reilly Media, Inc., 2008.

[7] S. Osinski and D. Weiss, "Conceptual clustering using lingo algorithm: Evaluation on open directory project data," in *In IIPWM04*, 2004, pp. 369–377.

[8] H. Marmanis and D. Babenko, *Algorithms of the Intelligent Web*. Greenwich, CT: Manning Publications Co., 2009.

[9] Global Terrorism Database, START, accessed on 9 December 2010. [Online]. Available: http://www.start.umd.edu/gtd/