# Creation of a Habit Model from GPS Data and Algorithms for Providing Awareness Services

Nobuo Matsuo
Graduate School of Engineering, Soka University
1-236 Tangi-cho, Hachioji-shi, 192-8577 Japan
e10m5235@soka.ac.jp

Kazumasa Takami
Graduate School of Engineering, Soka University
1-236 Tangi-cho, Hachioji-shi, 192-8577 Japan
k_takami@t.soka.ac.jp

***Abstract*—Big data including life logs, are attracting attention because of a number of recent developments: ever-increasing volume of data being generated every day due to advances in the broadband environment, increasing sophistication of mobile terminals, and growth of social networking services. If these can be fully exploited, useful added value can be created. Collecting, analyzing and managing such a huge volume of diverse data require innovative ideas and technologies. As an example of life log-based services, this paper proposes a "awareness" service, which provides a reminder for the user of a cellular phone or smartphone based on his/her particular situation (time, location, etc.) in order to encourage him/her to take a certain action. A convenient feature of life logs is that the user's location can be determined easily, even without requiring his/her conscious action, because the global positioning system (GPS) is normally installed in these mobile terminals. The algorithms for analyzing the user's GPS data to identify significant data in them, and for deriving his/her habit model from the identified data have been proposed. These algorithms have been implemented and evaluated. The awareness services have been compared with corresponding existing services, and the volume of data that can be reduced by using the habit data has been calculated.**

***Keywords—GPS; big data; location information; a habit model: awareness service***

## I. INTRODUCTION

Big data are attracting attention because of a number of recent developments: ever-increasing volume of data being generated every day due to advances in the broadband environment, increasing sophistication of mobile terminals, and growth of social networking services. Big data is a huge volume of diverse and indeterminate data, such as a life log, which records the activities of an individual. Useful added value can be created if big data is fully exploited. However, big data is so big, amounting to 1.8 trillion GB in 2011, and so diverse that it is not easy to link, analyze and manage big data. Innovative ideas and technologies are required to execute these. A life log that records the entire life of a person can be as big as 3 TB, and contains wide-ranging types of data. A technology to record life log data continuously must be devised. One of the main items of a life log is location information. This information can be obtained easily because the global positioning system (GPS) is normally installed in cellular phones and smartphones, and because the information is recorded even without requiring the user's conscious action. In this paper, we have limited the scope of a life log to GPS data, and studied how to analyze, manage and utilize GPS log data.

The goal of this research is to develop a life log service for cellular phone and smartphone users. We have proposed algorithms for a awareness service, a service that analyzes the user's GPS data to identify significant data in them, derives his/her habit model from the identified data, and provides a reminder for the user based on the habit model and his/her context (time, location, etc.). These algorithms have been implemented and evaluated. The awareness services have been compared with corresponding existing services, and the volume of data that can be reduced by using the habit data has been calculated.

Section II discusses related studies. Section III outlines the habit model used in this research and the proposed awareness services. Section IV identifies the issues that need to be addressed to develop this service. Section V provides algorithms that address these issues. Section VI evaluates the proposed algorithms and the proposed service. Finally, Section VII summarizes this paper, and presents issues that need to be studied.

## II. RELATED WORK

The scope of research on life logs encompasses the collection, analysis, management and utilization of life logs. Although the collection and analysis of life logs have been studied widely [1][2], there are few studies on the management and utilization of life logs [3][4]. Some studies on the collection and analysis of GPS log data focused on identifying user locations, and developing a model that indicates transitions of user locations. For example, Ashbrook et al. determined the user's location from GPS data using the k-means clustering, and proposed to predict the his/her future movement using a Markov model by assigning a probability to each possible transition from his/her current location [5].

Nishino et al. determined the user's location using the DBSCAN clustering, accumulated data in which the transitions of locations are sorted chronologically, and applied sequential mining to the accumulated data to develop a model that indicates transitions of the locations frequented by the

user [6]. Both studies determined the user's location and developed a model that indicates the transition from the determined location. They focused on the analysis of GPS data, but did not go further to study how to manage and utilize the analysis result. Neither did they consider the means and routes of movements, information that is related to transitions of the user's locations.

Our study differs from these studies in a number of respects. First, in our study, the granularity of locations is buildings (and their premises). We have not only developed a model for transitions of the user's location but also used the model for data management. Second, we consider not only the user's location and the duration of stay at a location but also other time data, such as day of the week, and the means and routes of movements, which constitute a part of the information about transitions of the user's location. Third, our study encompasses the entire scope of log data handling, from the collection to the management and utilization of log data.

## III. Habit Model And Awareness Service

### A. Habit model and Awareness

A habit model represents the user's behavior in the form of transitions of his/her locations. It is created by accumulating and synthesizing data on daily transitions of the user's locations. "Awareness" is meant in this paper to encourage the user to take an action that he/she is expected to take next, or to alert him or her to a certain event. Figure 1 shows a conceptual diagram of a habit model.

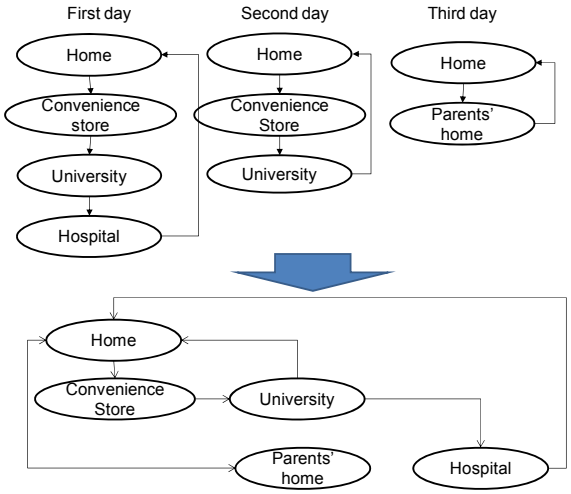



Fig. 1. Habit Model.

### B. Awareness service that is based on the habit model

In this paper, the awareness service, one of life log services, consists of the basic service, which provides a reminder, and a supplementary service, which offers information related to an action that the user may take based on the reminder. The types of awareness considered are the following:

- Awareness that informs the user of the approaching departure time of the last train or bus.
- Reminder that informs the user to the approaching deadline for returning DVDs to a rental shop or books to a library.
- Awareness that informs the user of a certain event (relocation of a shop, etc.)
- Awareness that helps the user avoids daily life risk.

## IV. Study Issues

The main issues that need to be addressed in developing the awareness service are as follows.

### A. How to Identify Significant Information in a GPS Log

A GPS log simply contains time, latitude and longitude data. It is necessary to study how to identify significant information in a GPS log and how to obtain it.

### B. How to Define and Derive a Habit Model

A habit model is used to predict the user's next action and to support him/her in taking that action. To define a habit model, it is necessary to study how information taken from a GPS log can be translated into an expression of action, and how a particular habit model can be derived. It is also necessary to study how to determine the direction of each transition from the user's location in the habit model in order to predict the user's next action.

### C. How to Provide the Awareness Service Based on the Habit Model

It is necessary to study the algorithms with which the reminder service can be provided based on the derived habit model and context (time, location, etc.) of the user.

## V. Solutions

### A. Experiment of Collecting a GPS Log and Findings

Humans repeat the cycle of moving, staying at a location, and moving again. We have collected a GPS log experimentally focusing on this cycle, analyzed it, and gained some insight into the speeds and directions of movements. Figure 2 shows the movements of a subject before and after he stopped at New Loire (cafeteria in Soka University), as plotted on the Google earth and the Google map.

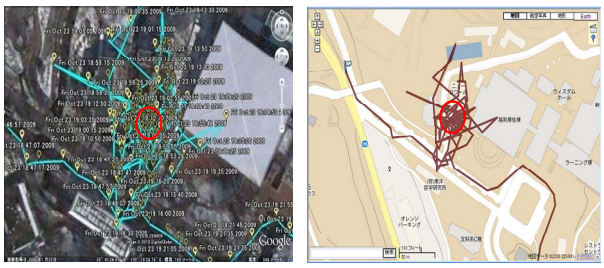


Fig. 2. GPS Log Obtained Using a GPS Logger.

*1) Moving Sections and Staying Sections:* Figure 3 shows changes in moving speed over time. The changes take a waveform that looks like a series of pulses. The higher levels

indicate the speed in moving sections, the sections in which the subject is moving from one building to another while the lower levels indicate the speed in staying sections, the sections in which the subject stays within a building (or its premises). Naturally, the speed is higher in moving sections than in

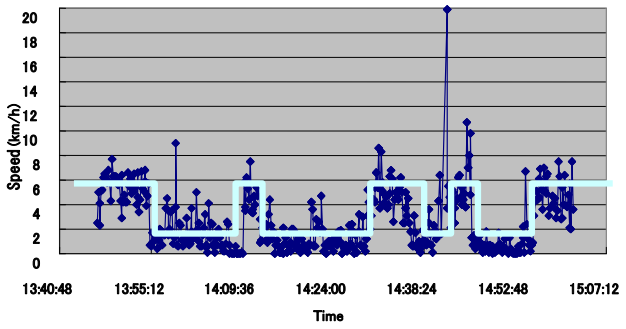



Fig. 3. Changes in moving speed over time.

*2) Findings about the directions of movements:* The difference in the latitude values between two points can be classified into plus, zero and minus. The difference in longitude values can also be classified in the same way. When these categories of differences in latitude and longitude values are combined, there can be 9 direction patterns as shown in Table I. Table II shows actual latitude and longitude values in both moving and staying sections recorded in the above-mentioned experiment. It can be observed that the movement directions are stable in moving sections, but unstable in staying sections.

TABLE I. CLASSIFICATION OF MOVEMENT DIRECTIONS

| 9 patterns of movement direction | | |
|---|---|---|
| Pattern name | Difference in latitude value | Difference in longitude value |
| A | + | + |
| B | + | - |
| C | + | 0 |
| D | - | - |
| E | - | + |
| F | - | 0 |
| G | 0 | 0 |
| H | 0 | + |
| I | 0 | - |

## B. Algorithms for Identifying Significant Information

*1) Significant Information to be Identified:* Information included in a GPS log can be broadly classified into three categories:

- Temporal information (date, day of the week, time)
- Staying building information (information about the building where the user is staying, such as building location and name, duration of stay, number of stays, etc.)
- Transition-related information (means, route, time and frequency of movement)

To identify staying building information and transition-related information, it is necessary to divide items of information in a GPS log into those in moving sections and those in staying sections. Staying building information can be found in information in staying sections while transition-related information can be found in information in moving sections. The algorithm for finding staying building information is described in this paper.

TABLE II. COMPARISON OF MOVEMENT PATTERNS BETWEEN THE MOVING AND STAYING SECTIONS

| Differences in latitude and longitude values in moving sections | | |
|---|---|---|
| Difference in latitude value | Difference in longitude value | Direction pattern |
| 0.00005720 | -0.00013730 | B |
| 0.00005340 | -0.00013730 | B |
| 0.00004200 | -0.00016790 | B |
| 0.00005340 | -0.00013730 | B |
| 0.00004950 | -0.00015260 | B |
| 0.00001910 | -0.00015260 | B |
| 0.00004580 | -0.00015260 | B |
| 0.00003060 | -0.00013730 | B |
| 0.00002290 | -0.00010680 | B |
| 0.00004580 | -0.00006110 | B |
| 0.00006110 | -0.00009160 | B |

| Differences in latitude and longitude values in staying sections | | |
|---|---|---|
| Difference in latitude value | Difference in longitude value | Direction pattern |
| -0.00004580 | 0.00000000 | F |
| 0.00002290 | 0.00000000 | A |
| 0.00003050 | 0.00003060 | A |
| 0.00002670 | -0.00003060 | B |
| 0.00005340 | 0.00007630 | A |
| 0.00006100 | 0.00000000 | C |
| 0.00004580 | -0.00007630 | B |
| -0.00001530 | -0.00001520 | D |
| -0.00011060 | -0.00003060 | D |
| -0.00018320 | 0.00016790 | E |
| 0.00003430 | 0.00000000 | A |

*2) Algorithm for Determining Moving and Staying Sections:* The algorithm for determining moving and staying sections has been derived based on the findings described in Section V.*A*. Figure 4 shows the proposed algorithm.

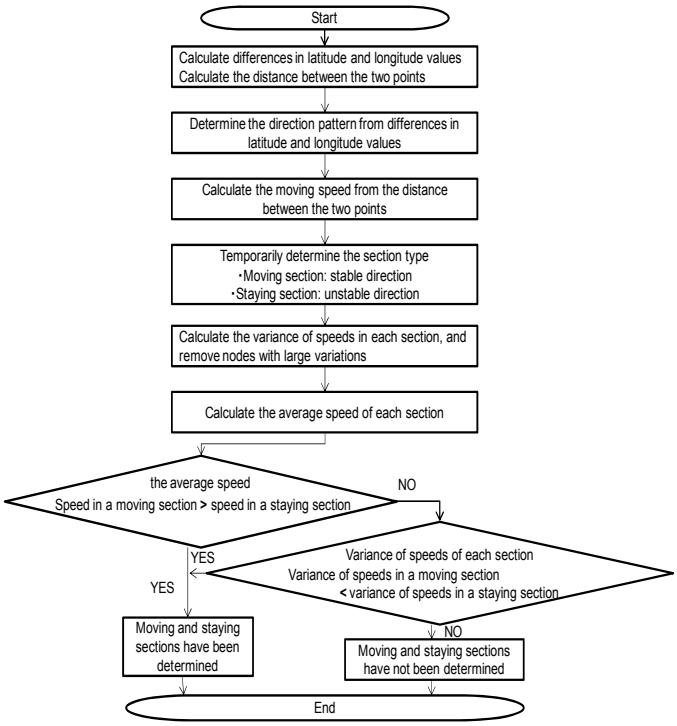



Fig. 4. Algorithm for Determining Moving and Staying Sections.

*3) Algorithms for Determining the Latitude and Longitude Values of a Staying section:* A staying section can contain several nodes, which are points in a building. Three alternative algorithms for selecting the optimal node whose location can be used to represent the location of the staying section are proposed below. In Alternative *A*, nodes with large errors are removed, and the average latitude and longitude values of the remaining nodes are used as the location of the section. In Alternative *B*, DBSCAN [7] is applied to the nodes to develop clusters, and the average latitude and longitude values of the nodes within the cluster that contains the nodes through which the user entered the building (starting node) and the node through which the user exited the building (ending node) are used as the location of the section. In Alternative *C*, the average latitude and longitude values of the nodes in the cluster in which the number of nodes with large errors is the smallest are used as the location of the section. These alternatives are shown in Figs. 5 to 7.

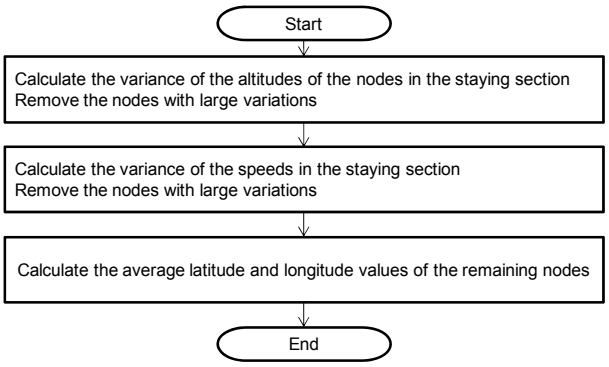



Fig. 5. Alternative A.

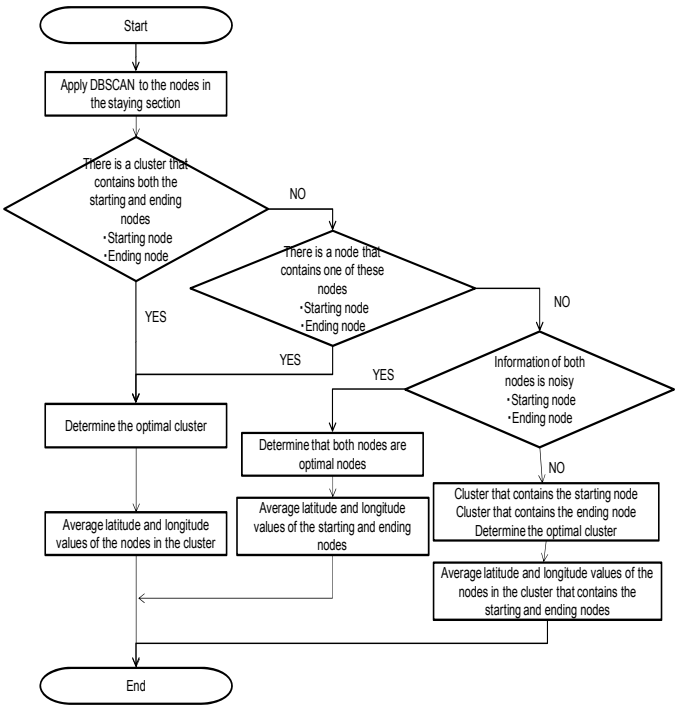



Fig. 6. Alternative B.

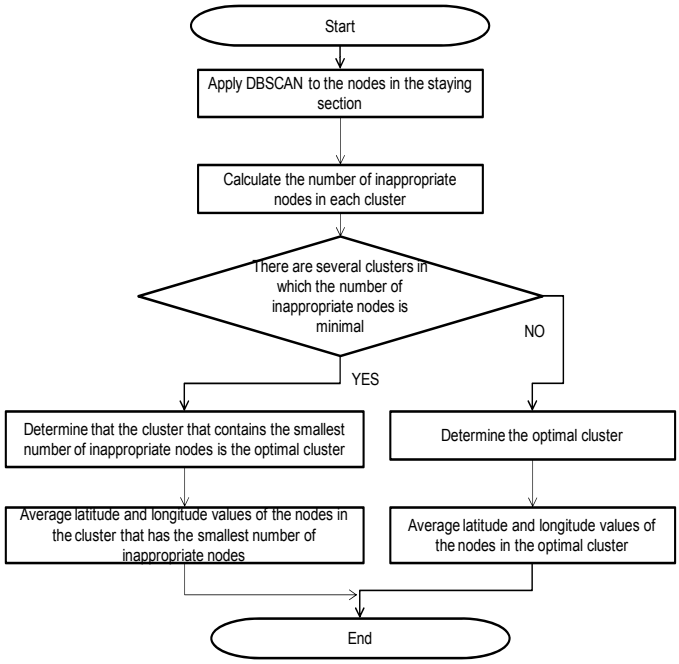



Fig. 7. Alternative C.

### C. *Definition of a Habit Model and Algorithm for Deriving the Habit Model*

*1) Definition of a Habit Model:* Human behavior may be represented by transitions of his/her locations. A habit model is defined as shown in Fig. 8 based on the locations where he/she stayed and the transitions from these locations.

As shown in Fig. 8, our habit model has a hierarchical structure. The upper layer consists of the original data while the lower layer stores processed data. Information about the location of stay are classified into STATE elements. Information about movement are classified into LINK elements. The elements that influence the selection of the destination are classified into TREE elements. "Log" stores all the processed data elements. The proposed awareness service is normally provided using only the information in the upper layer but information in the lower layer may be used as necessary.

*2) Definition of a Habit Model:* The habit model is derived as follows. First, Log-DB is built, followed by the construction of State-DB, Link-DB, and Tree-DB. New log data to be entered in a particular database is compared with data in the database, and if there is matching data in the database, the update frequency is changed. If there is no matching data, it is entered in the database. Duration of stay, etc., are obtained by referring to data in Log-DB. The steps for constructing State-DB, Link-DB, and Tree-DB are shown in Figs. 9, 10 and 11.

**Habit model layer**

STATE elements: stored in StateDB

| Location name | Representative GPS data | Day of the week | Time of the day of The starting time | Average starting time | Duration of stay | | | Number of stays | Date of last update |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Maximum duration of stay | Minimum duration of stay | Average duration of stay | | |

| LINK elements: stored in LinkDB | | | | | |
|---|---|---|---|---|---|
| Starting location | Destination | Frequency | Moving time | Means of movement | Name of the route of movement |

| TREE elements: stored in TreeDB | | | | | | |
|---|---|---|---|---|---|---|
| Starting location | Destination | Frequency | Day of the week | Time of the day | Means of movement | Name of the route of movement |

**Processed data layer**

GPS Log data (Processed data) : LogDB

| Starting location | Representative GPS data of the starting location | Destination | Representative GPS data of the destination | Starting time of stay at the starting location | Duration of stay at the starting location | Date | Day of the week | Means of movement | Moving time | Route of movement |
|---|---|---|---|---|---|---|---|---|---|---|

Fig. 8. Definition of a Habit Model.

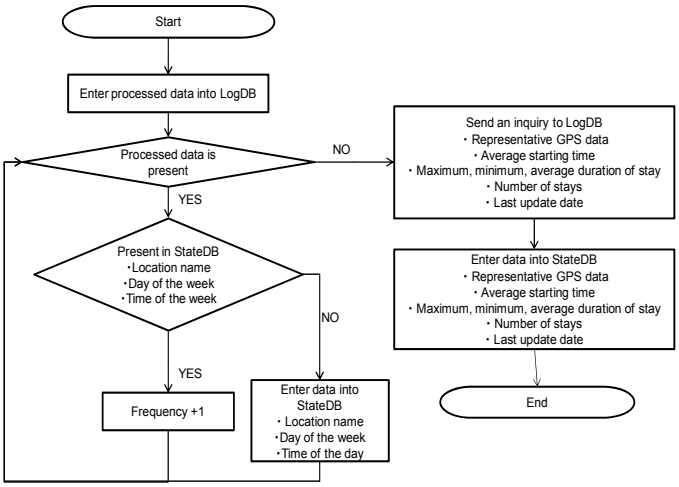



Fig. 9. Steps of constructing StateDB.

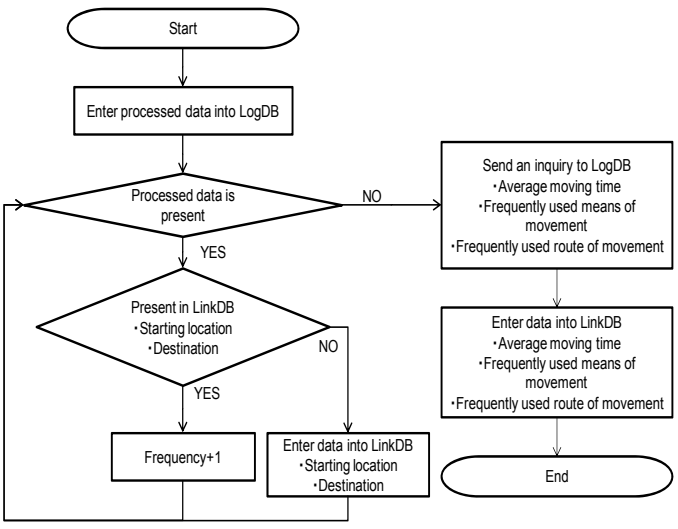



Fig. 10. Steps of constructing LinkDB.

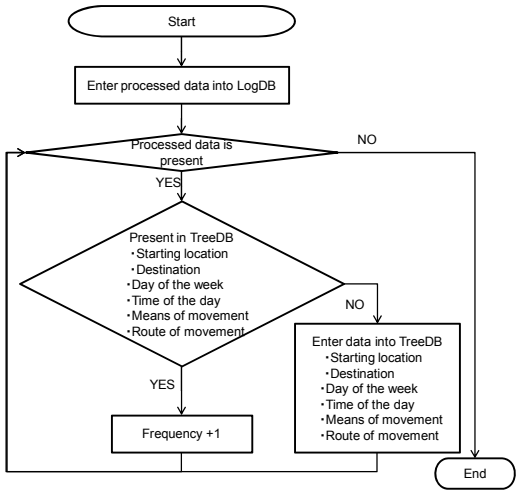



Fig. 11. Steps of constructing TreeDB.

*3) Determination of Destinations in the Habit Model:* Since this habit model is a transition model of locations, it resembles a state transition model. The destination of a transition can be determined based the current state of the user. In this paper, destinations are determined based on data in Tree-DB, which contains well-defined transition conditions derived from empirical data. Specifically, the next destination of the user is determined based on information about the day of the week, time of the day, means of movement available and route of movement. Data elements in Tree-DB are shown in Fig. 8. In Tree-DB, *destination* is a prediction result while *day of the week, time of the day, means of movement*, and *name of the route of movement* are transition conditions.

## *D. How to Provide the Awareness Service Using the Habit Model*

An example of awareness services that use the habit model is the provision of the information about the last train and bus. This service is provided as follows:

Step 1: Search State-DB for information about the building in which the user stays.

Step 2: Search Tree-DB to determine his/her destination.
Step 3: Search Link-DB to obtain information about the means of movement.
Step 4: Obtain information about the train station or bus stop the closest to the current location and that the closest to the user's home.
Step 5: Obtain information about the last train and bus based on the information obtained in Steps 1 to 4.
Step 6: Advice the user of the information about the last train and bus if the user still remains at the location identified in Step 1.

## VI. EVALUATION

### A. *Evaluation of the Algorithm for Identifying Significant Information*

*1) Evaluation of the algorithm for determining whether the user is moving or stays at a location:* An experiment was conducted in which the user visited 25 buildings in Hachioji, Tokyo. Log data were input in the log data analysis program, which contains the algorithm for determining whether the user is moving or stays at a location (within a building). Table III shows the evaluation result of this algorithm. The algorithm correctly determined that the user stayed at a location at 24 of the total of 25 staying sections (96%). The algorithm erred in one staying section because it failed to classify the log data of that section correctly into those with stable movements and those with unstable movements. The algorithm correctly determined that the user was moving in 38 of the total of 39 moving sections (97.4%).

TABLE III. EVALUATION OF THE ALGORITHM FOR DETERMINING WHETHER THE USER IS MOVING OR STAYS AT A LOCATION

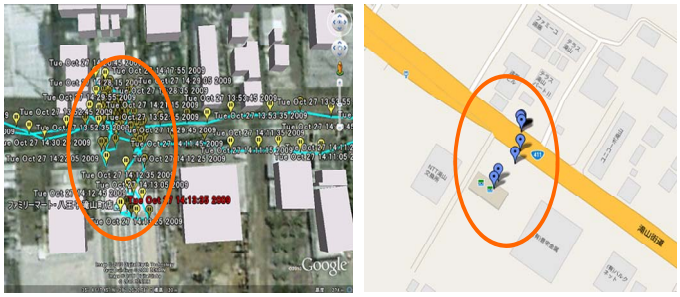
| Evaluation of the Algorithm for Determining whether the User is Moving or Stays at a Location | | |
|---|---|---|
| Criteria / Evaluation Item | Percentage of correct determinations | Number of the correctly determined sections/total number of sections |
| Determination of staying sections | 96.0% | 24/25 |
| Determination of moving sections | 97.4% | 38/39 |

*2) Evaluation of the algorithms for determining the latitude and longitude values:* Out of the 24 staying sections where the above algorithm correctly determined that these were staying sections, two sections in which Google map had not been updated were excluded in the evaluation of the algorithms for determining the latitude and longitude values. Table IV shows the result of the evaluation of the three alternative algorithms described in Section *V.B.3)*.

Table IV reveals the following. Out of the 22 staying sections, 7 positions determined by Alternative *A* correctly fell within the respective buildings. The percentage of correct positions was 31.8%. If an error of 10 m is tolerated, 16 positions were correctly determined (72.7%). With Alternative *B*, 6 positions were within the respective buildings (27%). If we allow an error of 10 m, the algorithm determined 14 positions correctly (63.6%). Alternative *C* determined 7 positions within the respective buildings (31.8%). With the tolerance of an error of 10 m, the number of correctly determined positions rose to 14 (63.6%). These results are not really satisfactory. However, we have confirmed that clustering ran correctly with Alternatives *B* and *C* as is shown in Fig. 12. Figure 12 (a) shows the state before clustering, and Fig. 12 (b) the average positions of the generated clusters. If the most appropriate cluster can be selected, the accuracy of determining the correct positions can be improved dramatically. We expect that even the above-proposed three alternative algorithms can prove effective if the algorithm and clustering the most appropriate for each section can be selected because this would raise the percentage of determining the positions correctly within the respective buildings to 50%, and even to about 80% if an error of 10 m is tolerated.

TABLE IV. EVALUATION OF THE ALGORITHMS FOR DETERMINING THE LATITUDE AND LONGITUDE VALUES

| Evaluation of the Algorithms for Determining the Latitude and Longitude values | | |
|---|---|---|
| | Alternative A | |
| Aalternative / Evaluation Item | Percentage of correct determinations of positions | Number of buildings for which the positions are determined correctly/the number of buildings |
| Percentage of correct determinations of positions within the respective buildings | 31.8% | 7/22 |
| When an error of 10 m is tolerated | 72.7% | 16/22 |
| | Alternative B | |
| Aalternative / Evaluation Item | Percentage of correct determinations of positions | Number of buildings for which the positions are determined correctly/the number of buildings |
| Percentage of correct determinations of positions within the respective buildings | 27.0% | 6/22 |
| When an error of 10 m is tolerated | 63.6% | 14/22 |
| | Alternative C | |
| Aalternative / Evaluation Item | Percentage of correct determinations of positions | Number of buildings for which the positions are determined correctly/the number of buildings |
| Percentage of correct determinations of positions within the respective buildings | 31.8% | 7/22 |
| When an error of 10 m is tolerated | 63.6% | 14/22 |



(a) GPS data before clustering (b) GPS data after clustering

Fig. 12. Result of clustering.

### B. *Evaluation of the Algorithm for Identifying Significant Information*

*1) Evaluation of the Effectiveness of the Habit Model:* To evaluate how effective the proposed habit model for the awareness service is, we have compared it with two corresponding existing services shown in Table V. The functions provided by the three services are compared in Table VI.

TABLE V. FUNCTIONS PROVIDED BY CORRESPONDING EXISTING SERVICES

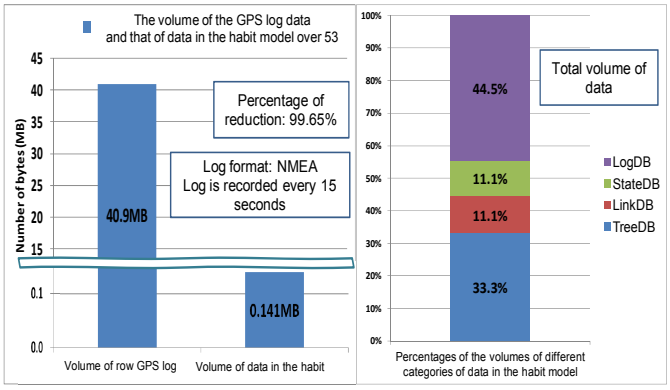
| Name of existing service | Functions provided |
|---|---|
| i-concierge [8] automatic GPS function (last train alarm) | The user pre-registers up to three destination stations. This service informs him/her of the departure time of the last train from the nearby station to the destination station at the time when he/she can still walk to the nearby station before the departure time. |
| i-concierge [8] automatic GPS function (bus operation state) | The user pre-registers the bus stop he/she uses to go to work or school. This service informs him/her of the arrival time of the next bus and that of the last bus at the bus stop. |

TABLE VI. COMPARISON WITH EXISTING SERVICES

| | i-concierge (last train) | i-concierge (last bus) | Proposed habit model |
|---|---|---|---|
| Initialization | Needed | Needed | Not needed |
| Scope of service | From the current location to home | Pre-registered bus line | From the current building to home or parent's home |
| Provision of linked information | Information about the last train only | Information about the last bus only | It is possible to link the information about the last train to information about the last bus |

To confirm that the habit model shown in Table VI is indeed more desirable than the existing services, we have implemented a part of the last train informing service. Since the habit model contains information about means of movements, it has been possible to link the information about the last train to information about the last bus.

*2) Evaluation of how much the use of the habit model can reduce the volume of data that has to be managed:* In using life logs, it is important to reduce their sizes as much as possible. Since the habit model manages data that have been processed from row GPS log data, the volume of data it manages is much smaller than that of the row data. The comparison of the volume of row GPS log data and the volume of data in the habit model is shown in Fig.13.




(a) Reduction in the volume of data from

(b)Percentages of the volumes of different categories of data in the habit model

Fig. 13. Reduction the Volume of Data.

Using a GPS Logger (tripmate850), we have collected GPS data every 15 seconds over 53 days from the middle of March to early April and from early September to the middle of November. The format of the collected data was that of NMEA (National Marine Electronics Association). Figure 13 (a) shows that the use of the habit model dramatically reduces the volume of data that has to be managed, from 40.9 MB of the row GPS data to 0.141 MB of the habit model -- a reduction of 99.65%. This reinforces the effectiveness of using the habit model. Figure 13 (b) shows the percentage of the size of each database in the habit model (Log-DB, State-DB, Link-DB, and Tree-DB). The largest database is Log-DB. Since it stores all processed data, its size will grow as the collection of GPS log continues. In contrast, other databases, State-DB, Link-DB, and Tree-DB, only update the same items of data. The volume of such update data tends to decrease as the collection of GPS log continues. Therefore, the proportion of Log-DB will continue to increase.

## VII. CONCLUSIONS

We have identified significant information in GPS data, built a habit model from this information, and proposed a awareness service using this habit model. An experimental system for providing this service has been developed, and used to compare this service with corresponding existing services. The comparison result has confirmed the effectiveness of the habit model. The use of the habit model also reduces the volume of data that has to be managed.

The issues that need to be studied include improvement of the algorithm for identifying various types of significant information, a service that makes use of the habit models of multiple persons, and possible services that can be provided by linking not only GPS data but also other types of data in life logs.

## REFERENCES


[1] Kazuya Okada, Teruaki Yokoyama, Youki Kadobayashi, and Suguru Yamaguchi , "Estimating Location Relationship from User Trajectories and its Application," pp1120-1126, DICOMO symposium 2009.7.

[2] Naoharu Yamada, Yoshinori Isoda, Masateru Minami, and Hiroyuki Morikawa, "Movement Prediction based on Route History for GPS-enabled Cellular Phones," B-15-23 IEICE General Conference 2010.3.

[3] Daisuke Kamisaka, Takeshi Iwamoto, and Hiroyuki Yokoyama, "Estimation Method of Homes and Offices from Location Data Obtained by Mobile Phones, " B-20-29 IEICE General Conference 2010.3.

[4] Manabu Motegi, Hirohisa Tezuka, Yukihiro Nakamura, Shin-ichiro Eitoku, Shun-ichi Seko, Masaaki Nishino, Shin-yo Muto, and Masanobu Abe, "Field Trial of the System using Life Logs," D-9-12 IEICE General Conference 2010.3 .

[5] Ashbrook D and Starner T, "Using GPS to learn significant locations and predict movement across multiple users," Personal and Ubiquitous Computing, Vol.7, NO.5, pp.275-286, 2003.

[6] Nishino Masaaki, Shunichi Seko, Masakatsu Aoki, Tomohiro Yamada, Shinyo Muto, and Masanobu Abe, "A Study on Extracting Movement Patterns from Transition Data," IPSJ SIG Technical Report UBI 2008(110),pp.57-64,2008.11.

[7] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise, " KDD'96,1996.

[8] i-concierge, http://www.nttdocomo.co.jp/service/customize/iconcier/, (2013/3/10)