Mining Life-log Sensing Data to Extract User's Significant Locations and Movement

NIKEN TRI MAHAYANI† TAKUMI KITAZAWA††
NISHIO NOBUHIKO††

Abstract: The expansion of location-acquisition technologies facilitates a log of user locations and trajectories. In the research, we applied truncating into the GPS log data to mine a list of user possibly significant locations and movements among them. The experiment result shows that truncating can be used for clustering alternative with benefit no need to determine the initial number of cluster .

Keywords: Data Mining, Life-log, Life-log, GPS

1. Introduction

The popularity of mobile phone that comes with sensor devices, such as an accelerometer and GPS, provides facility logging of locations and movement histories of a user. Afterward the user's locations and trajectories can be tracked using such devices.

By aggregating and analyzing GPS log of the user, a lot of useful information can be achieved such as a list of significant locations of a user. Then, we can use the accumulating knowledge of significant locations from each user to find out the frequently visited locations by those users. Henceforth, we can rank the locations and share this information among the users. Further, a user can find some locations based on the reference knowledge of other users. In other word, a location recommender system can be built upon the system.

However, the recommender system based only on the knowledge of the user's significant locations is not adequate to give proper recommendation. The transportation used by the user has to be considered as important role to decide which location should be recommended to the user. For example, if we discover a user that is walking, the recommender system cannot recommend a restaurant that has distance more than five kilometers from the user located.

In addition, both the information of the user's significant locations and the user's transportation modes also can be embedded into another useful system, for instance, precise advertising, and route arrangement. Realize that the user location and transportation awareness can be very useful become the reason of choosing 'Mining Life-log Sensing Data to Extract User's Significant Locations and Movement' as our research focus.

The drafts' contributions are two parts. First, a list of a user's significant locations is extracted. We use the unavailability of GPS signal to penetrate into the building as evidence that the user has gone indoors. It is also a common sense that the user speed in their significant locations will not exceed the walking speed. The duration during the GPS signal lost, the speed shortly before and after the GPS signal lost are used as thresholds to extract the user's significant location candidates from the raw GPS data. Due to the fact that GPS log with different coordinates may represent the same location, the clustering algorithm is needed to overcome the problem appeared. Therefore, truncating is applied to the significant location candidates. By truncating, the GPS points that refer to the same location will be group into single GPS point. Khetarpaul, S. et al. [4] stated truncating to three decimal places corresponds to a physical distance of about 100 meters. Second, we address the user's transportation mode detection using an accelerometer data and the speed of the user. The accelerometer data analyzed can be used to differentiate whether the user in standing position or walking position. After the user's significant locations are extracted, the trajectories movements among the locations are analyzed. In the analyzing process, the trajectories are divided into segments based on the accelerometer data walking result. From the speed of each segment, the features of the transportation modes are identified. In the next phase, the user's transportation mode can be inferred.

The rest of this draft is organized as follows. Section 2 discusses the existing studies about mining significant locations and mode transportations detection. Section 3 discusses our proposed solution. Section 4 concludes our approach result.

2. Related Work

There has been a lot of prior work on analyzing raw GPS data to detect user significant locations. In "CityVoyager", Takeuchi, Y. et al. [1] use the unavailability of GPS signals as evidence that the user has gone indoors. The system records the user's shop visiting histories based on GPS log and gives some shops recommendations which are similar to user's previously visited shops. The system uses an item-based collaborative filtering method. Ashbrook, D. et al. [2] use K-mean clustering of visited places to find frequented places, where visited places are defined as places where GPS signals were continuously lost, or places where user movement was slower than one mile per hour. In our approach, we keep the usage of unavailability of GPS signals in indoor area as evidence, but we replace the usage of K-mean clustering with truncating as another alternative for clustering. As truncating advantage, determining the initial number of clusters is not required. Zhen, Yu. et al. [3] use a tree-based hierarchical graph (TBHG) to model multiple users'

† Graduate School of Computer Science and Engineering, Ritsumeikan University
†† Ritsumeikan University

travel sequences on a variety of geospatial scales based on GPS trajectories. Khetarpaul, S. et al. [4] analyzed and mined GPS trajectories of multiple users then applied various relational algebra and statistical operations to find interesting locations. Various operations like bag union, mode are applied to obtain a ranked list of interesting locations. Zhen, Yu. et al. [5] is based on transportation mode detection from only GPS data and the result is to distinguish the walking mode with non-walking mode.

3. Proposed Solution

In this section, we first define some terms used in this draft and then briefly describe procedures of extracting significant locations and detecting mode of transportation. Our first approach is for extracting significant locations consist of a filtering phase and a truncating phase. The second approach is for detecting transportation modes consist of a segmentation method and a feature extraction.

3.1 Preliminary

In this subsection, we define some terms used in our approach as follows.

- **Term 1. GPS Log (*L*)** is an information record of geographical areas passed by the user sequences in time $L = \{l_1, l_2, \ldots, l_n\}$ $\forall\, l_i \in L$ consist of user identity (u_i), latitude $(l_i.Lat)$, longitude $(l_i.Lon)$, velocity $(l_i.Vel)$, accuracy $(l_i.Acc)$ and timestamp $(l_i.Ts)$.
- **Term 2. GPS trajectory (*LTraj*)** is a connection among GPS points in their sequential time $LTraj = l_1 \rightarrow l_2 \rightarrow \ldots \rightarrow l_n$, where $l_i \in L$, $l_i{+}1.Time > l_i.Time$.
- **Term 3. Significant location candidate (*Sc*)** is a collection of GPS points that qualify a duration threshold *(ThreshDuration)*, a speed threshold *(ThreshSpeed)* and an accuracy threshold *(ThreshAccuracy)*. $Sc = \{ sc_1, sc_2, \ldots, sc_n \}$, where
  $sc_n.TimeGPSfound - sc_i.TimeGPSlost \geq ThreshDuration;$
  $sc_n.SpeedGPSfound \,\&\, sc_i.SpeedGPSlost \leq ThreshSpeed;$
  $sc_n.AccuracyGPS \leq ThreshAccuracy.$
  The significant location candidates will be occurred with three conditions:
  1. The duration from a user loss GPS satellite signal until get back the GPS satellite signal equal or more than a duration threshold. The duration threshold is the minimum period of GPS time lost required to detect a significant location candidate. In most cases, this happen when the user has gone indoors and the building construction barricade the GPS signals.
  2. The speed shortly before and after the GPS signal lost is equal or less than a speed threshold.
  3. The GPS accuracy is equal or less than an accuracy threshold.
- **Term 4. Significant location (*S*)**: A significant location *S* stands for a geographic region where a user stayed over a certain time period. The significant locations are formed from clustering the significant location candidates found. $S = \{ s_1, s_2, \ldots, s_n \} \forall\, s_i \in Sc,\ s_{i+1}.Time > s_{i+1}.Time.$
  $s_i.Time = \sum_{i=1}^{n} sc_1.Time + sc_2.Time + \cdots sc_n.Time$ where all sc_i in the same cluster and ordered consecutively.
- **Term 5. Truncating (*TruncCluster*)** is eliminating each GPS point of the significant location candidates detected into three digits decimal. The truncating into three digits decimal will give us physical distance around 100 meters (Khetarpaul, S. et al. [4]). In addition, all the GPS points around 100 meters will aggregate into only single cluster. We will assume the single cluster as a cluster of user significant location. From all GPS points in each cluster, we choose the GPS point with the high accuracy as significant location representation. Our terms can be described by Figure 1 as follows.

	UserId,	Latitude,	Longitude,	Accuracy,	Velocity,	Timestamp,
l1	u1	Lat1	Lon1	Acc1	Vel1	Ts1
l2	u2	Lat2	Lon2	Acc2	Vel2	Ts2
ln	un	Latn	Lonn	Accn	Veln	Tsn

A GPS log

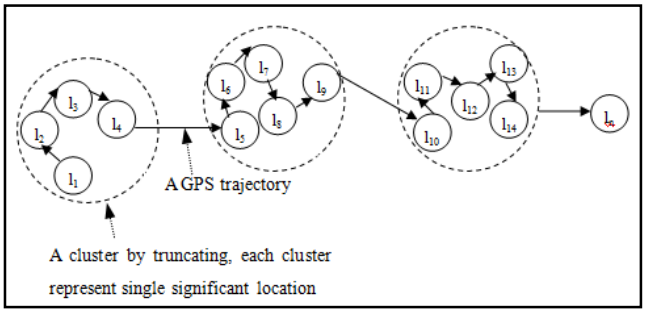

Figure 1. A GPS log, a GPS trajectory and three significant locations

3.2 Extracting significant locations

The process for extraction significant locations consists of two phases that are filtering phase and truncating phase.

3.2.1 Filtering phase

In filtering phase, the significant location candidates are filtered out from the raw GPS data. For the filtering process, three thresholds are applied. The thresholds are threshold duration (*ThresDuration*), threshold speed (*ThreshSpeed*) and threshold accuracy (*ThreshAccuracy*). For deciding appropriate threshold duration, we summarize the number of significant locations candidates found using various durations in Figure 2.

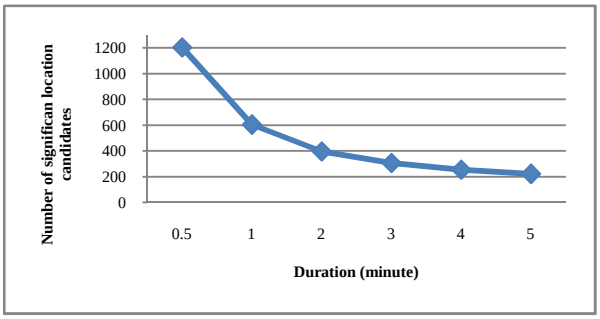

Figure 2. Number of significant location candidates found using various durations

As expected, the shorter threshold duration is chosen, the number of significant location candidates found will be increase.

Unfortunately, based only on the Figure 2 is difficult to decide appropriate threshold duration. In our application, 1 minute is chosen as threshold duration based on two reasons, which are:

1. The locations which the users only need to spend a moment for their activities, such as pick up laundry, want to be included.
2. Choosing threshold duration less than 1 minute will increase the amount of noise and the higher threshold duration will increase the possibility of some significant locations neglected.

For threshold speed is determined to 5 KM/H due to the average human walking speed and threshold accuracy is set to 64 as default value but the user can input another threshold accuracy value if it is desired. Algorithm filtering below gives a formal description of this process.

Algorithm Filtering

Input : GPS log raw data, L = $\{l_i\}$;
$l_i = \{u_i, lat_i, lng_i, Vel_i, Acc_i, Ts_i\}$

Output : Significant location candidates, Sc = $\{sc_i\}$
$sc_i = \{u_i, lat_i, lng_i, Ts_{i+1}, Ts_i\}$

1. **Begin**
2. **for** each GPS log raw data L **do**
3. duration <- (Ts_{i+1} - Ts_i)
4. distance <- *DistanceHubeny*
5. speed <- distance / duration
6. **if** (Accuracy < = *ThresAccuracy*) &&
7. (duration >= *ThresDuration*) **then**
8. **if** (speed.Ts_{i+1} <= *ThresSpeed) &&*
9. (speed.Ts_i <= *ThresSpeed)* **then**
10. Sc <- L
11. **End**

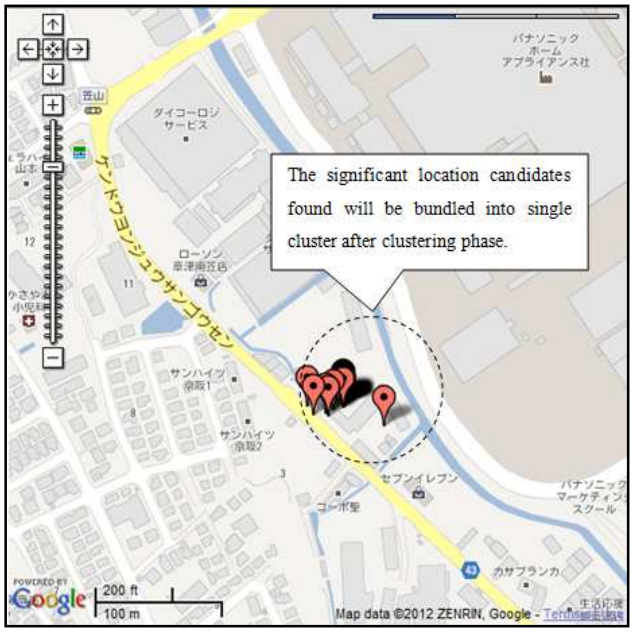

Figure 3. Example of the significant location candidates from a user during one day GPS data

Figure 3 shows the example result from the algorithm filtering. In this example, each marker represents one significant location candidate. As stated before that GPS log with different coordinates may represent the same location, all the markers must be checked whether they are refer to the same location. The following truncating phase presents to overcome the problem.

3.2.2 Truncating phase

After filtering out significant location candidates from raw GPS log data, all the gps points from the significant location candidates will be truncated into three digits decimal. The truncating GPS points will be assumed as clusters. Thus, all GPS points which are have the same truncating value, in other word they have the same cluster, will be treated as one single significant location. Then, the GPS point which has the highest accuracy in each cluster will be chosen as a significant location representation.

Figure 4. Different GPS coordinates represent the same location

	Distance (meter)	GPS Coordinate 1	2	3	Truncating into 3 digits
GPS	**Coordinate 1**	0	5.37	37.60	34.983, 135.950
	Coordinate 2	5.37	0	34.91	34.983, 135.950
	Coordinate 3	37.60	34.91	0	34.983, 135.950

** Haversine formula is used for calculation*

***The distance calculation results are rounded into two digits decimal*

Table 1. Distance calculation among coordinates in the same cluster created by truncating into 3 digits

Figure 4 shows three different GPS coordinates which are GPS coordinate 1 (34.9838376, 135.9500653), GPS coordinate 2 (34.98378932, 135.9500653) and GPS coordinate 3 (34.98364985, 135.9504086). These three GPS coordinates give the same truncating result that is (34.983, 135.950). Therefore, all the three GPS coordinates will be clustered into the same cluster. In other word, the three different GPS coordinates represent the same location. Table 1 shows the calculation distance among the GPS coordinates. Algorithm truncating

below gives a formal description of this process.

Algorithm Truncating

Input : Significant location candidates, Sc = { sc_i} ;
sc_i = { u_i, lat_i, lng_i, Ts_{i+1}, Ts_i}

Output : Significant location, S = { S_i}
S_i = {TruncCluster, u_i, lat_i, lng_i, Ts_{i+1}, Ts_i, $Stay_i$}

Begin

1. **for** each Sc **do**
2. TruncCluster <- 3 digits truncating of Sc
3. **for** each TruncCluster **do**
4. **while** TruncCluster **isSame do**
5. $Stay_i$ <- Ts_n - Ts_i
6. S_i <- sc_i
7. **End**

Figure 5. Example of the significant location from significant location candidates depicted in Figure 3.

Figure 5 shows the significant location resulted by truncating of the significant location candidates depicted in Figure 3.

3.3 Transportation mode detection

For detecting transportation mode, the trajectories among significant locations will be divided into many segments based on walking mode detected along the trajectories. Thus, walking mode detection will be done before creating the segments. After the segments created, the features from each segment will be extracted. The next step is inferring transportation modes. The transportation modes that will be inferred are *Walking mode*, *Bicycle mode*, *Bus mode*, *Private car mode*, *Taxi mode* and *Train mode*. The flow chart of the transportation mode detection is depicted in Figure 6.

The following two subsections will describe our approach for creating segments and extracting features.

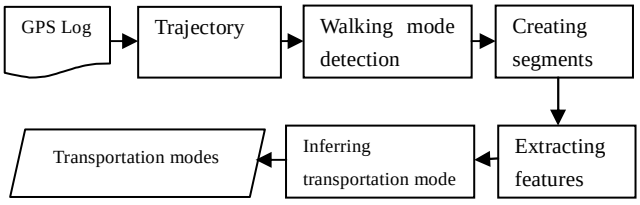

Figure 6. The flowchart of inferring transportation modes

3.3.1 Creating segments

The creating segments will consider the common fact that for changing from one transportation mode into another, people firstly must stop from their first transportation then go to the next transportation. In another word, the speed of the people will be change from the speed of transportation mode into the speed of human walking. Hence, the walking mode will be primary guidance to detect a transition between two transportation modes.

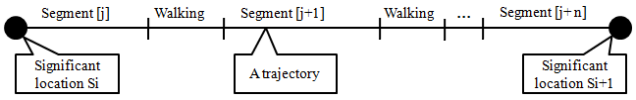

Figure 7. The segmentation based on walking mode.

Therefore, the creating segments should be based on walking mode detection. The new segment will be created if walking mode is detected. The walking mode detection can be done by analyzing the speed resulted from GPS data or/and by analyzing the accelerometer data. In our future research, we will use both the analyzing speed from GPS data and analyzing the accelerometer data with reason only using the GPS data will lead to some errors walking mode detection. The analyzing speed from GPS data are highly affected by traffic and weather condition. It can be easily understandable that the speed from bus mode will be go down as bicycle speed then walking speed during traffic jam. This condition will lead to the error walking mode detection. Hence, the accelerometer comes to overcome the problem. The accelerometer data analysis can differ whether user moving or staying. The walking mode will not be inferred if the GPS data infer the walking mode but the accelerometer data infer the user is staying.

3.3.2 Extracting features

The extracting features of each transportation mode will use inputs as follows:

1. Both the average speed and the maximum speed from each segment will be explored to extract the features of transportation modes. The maximum speed from the segments will be compared with the maximum speed boundary of each transportation mode. In the future, the maximum speed boundary of each transportation mode will be investigated from the GPS log data collected.
2. To differentiate between *Bus mode*, *Private car mode* and *Taxi mode* will use the information from the location where each mode of the transportations stop. For example, if the location where the transportation stop is identified as bus stop then the transportation will have more probability as *Bus mode* than *Private car mode* or *Taxi mode*. As well as if the location where the transportation stop is identified as parking user's own car then the transportation will have

more probability as *Private car mode* than others. Hence, the location labeling will be manually conducted at the locations where the transportations stop. For example, if the transportation stop is identified as a bus stop, the bus stop label will be given to that location.

3. The other inputs that can be used for the distinction between *Bus mode, Private car mode* and *Taxi mode* are the number of stop and the distances among the stops. For example, the transportation that has some stops and the distances among the stops are not so far will be more considered as *Bus mode*.

4. Conclusion and Future Work

Here, we conclude that extract user's significant locations can be done by GPS raw data analysis and many potential applications can be built upon the knowledge. Truncating can be used for another clustering solution in order to extract user's significant location.

In the future, we will implement the segmentation based on walking detection method and will apply the feature extraction approach followed by the result from our approach. We will also seek an appropriate method for inferring the transportation modes. In the first trial, we will try Decision Tree for inferring model.

Reference

1) Takeuchi, Y. and Sugimoto, M. CityVoyager: An Outdoor Recommendation System Based on User Location History. In Proceedings of UIC'2006, (Berlin, 2006), Springer Press: 625-636.
2) Ashbrook, D. and Starner, T. Learning Significant Locations and Predicting User Movement with GPS. In Proc. of 6th IEEE Intl. Symp. on Wearable Computers, 2002.
3) Zheng, Yu., Zhang, Lizhu., Xie, Xing. and Ma, Wei-Ying. Mining Interesting Locations and Travel Sequences from GPS Trajectories.
4) Khetarpaul, S., Chauhan, R., K Gupta, S., Subramaniam, V. and Nambiar, U. Mining GPS Data to Determine Interesting Locations. ACM 2011.
5) Zheng, Yu., Liu, Like., Wang, Longhao., Xie, Xing. Learning Transportation Mode from Raw GPS Data for Geographic Application on the Web. ACM 2008.
6) Liao, L., J Patterson, D., Fox, D. and Kautz, Z. Building Personal Maps from GPS Data. In Proceedings of the National Conference on Artificial Intelligent. ACM Press, 2004.