International Journal of Database Theory and Application
Vol.8, No.4 (2015), pp.149-156
http://dx.doi.org/10.14257/ijdta.2015.8.4.15

Mining Frequent Spatio-Temporal Items in Trajectory Data

Fengjiao Yin, Xu Li, Chunlong Yao, Lan Shen

School of Information Science and Engineering
Dalian Polytechnic University
No.1, Qinggongyuan, Ganjingzi District, Dalian 116034, P. R. China
lixu@dlpu.edu.cn

Abstract

The time aspect is not currently taken into account for finding a region of interesting (ROI) or a hot region, so that due to the time to visit frequently a place cannot be determined, it is difficult to discover the visiting regularity for a moving object. To this end, the spatio-temporal item (STI) and frequent spatio-temporal item (FSTI) integrated spatial and temporal attributes are defined. The FSTIs can represent a moving object often visits which area in what time, which can provide more useful information to improve the level of the location-based services(LBS). In order to find FSTIs, STIs are generated by using a density-based clustering algorithm to recognize the stay regions of objects, and then the STIs are mapped to 3D-grids integrated spatial and temporal dimensions. Finally, the extraction - merger strategy is used on the frequent grid cells to recombine the FSTIs. Experimental results on real dataset show that the approach proposed for mining FSTIs is effective.

Keywords: *Data Mining; Trajectory Data; Frequent spatio-temporal Items; 3D-grids*

1. Introduction

Global positioning systems (GPS) have become increasingly available and a large amount of spatio-temporal data is being generated. By analyzing the historical trajectories of users over a long time, we can find out the regularity behaviors of users, which are potential support for the decision-making in the future. For some intelligent location-based services, e.g., an intelligent ridesharing application, if some users tend to stay in the workplace almost every weekday approximately during the same time, vehicles can be shared by them and some important notes or location-based advertisings can be given to them instantly.

A region of interesting (ROI) or a hot region is proposed to indicate a spatial region, which is visited frequently by moving objects and has attracted increasing attention. K. E. Liu et al. [1] introduced a grid-based approximate schema to construction the dense regions. Y. Liu et al. [2] introduced improved density-based Clustering Ⅱ algorithm to mine all of the hot regions on different granularities based on the stay point sets. Because these methods focus on spatial frequency and ignore the temporal information, the frequent items are mentioned. It is obvious that the classical methods like Apriori [3] and FP-growth algorithm [4] cannot be directly applied to mining frequent items. Because the trajectory data is a temporally ordered sequence and contains complex semantics information due to mixture of the temporal and spatial relationships. [5]

Some item-like definitions have been proposed. L.Wang et al. [5] defined the item as an *Information Pair* [sid_i, $end\text{-}time_i$, sup_i] where sid_i is trajectory identifier, *end-time* is corresponding timestamp in the trajectory sid_i and *supi* is support value. L.Chen et al. [6] defined the item as a couple S_i = (R_i, T_i) where region R_i (i = 0, ..., k) is a set of merged neighboring cells, and T_i (i = 0, …, k) = ($T_{in}^{(i)}$, $T_{out}^{(i)}$), $\forall_{0\le i<k} T_{in}^{(i)} < T_{out}^{(i)}$, $T_{out}^{(i)} \leqslant T_{in}^{(i+1)}$. R_i is the

ISSN: 2005-4270 IJDTA
Copyright ⓒ 2015 SERSC

*i*th ROI. $T_{in}^{(i)}$ and $T_{out}^{(i)}$ are the times the user entering and leaving R_i respectively. Given two couples S_n and S_m (m not equal to n) in a Regional-Temporal Sequence (RTS), although R_n may equal R_m (as the user may revisit the same ROI in a single trip), T_n never equals T_m. In other words, the spatial component of a RTS may repeat, but the temporal component always increases. The entering time $T_{in}^{(i)}$ of region R_i is set as the time stamp of the first position in the region, while the leaving time $T_{out}^{(i)}$ is set as the last time stamp in that region. Although this definition of item can reflect the aforementioned regularity, it is just as a medium to describe the patent. In [7], the item is defined by the square cells and a five minutes interval centered around time instances written inside the square.

Although there are many works [7-9] focusing on the frequent spatial-temporal information, they cannot reflect a moving object often visits which area in what time. We give the definitions of spatio-temporal item (STI) and frequent spatio-temporal item (FSTI) integrated temporal and temporal attributes to substitute for this regularity. Correspondingly we give the approach for mining FSTIs. Step1, stay regions are recognized through a self-adaptive method [10] and STIs are generated by using a density-based clustering algorithm. Step2, in order to integrate the spatio-temporal information, the STIs will be mapped to 3D-grids. Step3, frequent grid cells are extracted and merged to recombine the FSTIs.

The remainder of the paper is organized as follows: Section 2 describes the generation process of STIs. Section 3 describes the Map algorithm in detail. Section 4 presents the Extraction - merger strategy. Section 5 illustrates the results and performances of this method. The last section concludes our approach and future work.

2. STIs Generation

2.1. Definition of STI

A trajectory is a temporally ordered sequence which record the spatial-temporal information of user, *Trajectory*= {P_1, P_2, ..., P_n}. Every point is composed of triples which contain latitude (*lat*), longitude (*lng*) and timestamp (*t*), P_i= (*lat*,*lng*,*t*) ($1 \le i \le n$).

Definition 1 (STI). Item *I*= {*Id*, *User_id*, *Stay_ region*, *Time_ interval*}, *Id* is a unique identifier of item. *User_id* is a unique identifier of user. *Stay_ region* is a geographic region that the user stays and *Time_ interval* is a period of time at this stay region.

Definition 2 (Stay region). *Stay_ region*={P_1, P_2, …, P_m }, $\forall P_i$=(*lat*, *lng*, *t*)($1 \le i \le m$), let $1 \le l \le m$, temporal threshold is α and spatial threshold is β. $\forall\Delta(P_i.\ t,\ P_l.t) \ge \alpha$, $\forall\Delta dist(P_i, P_l) \le \beta$; each $\Delta(P_i.\ t,\ P_l.\ t)$ is the time interval between P_i and P_l. Each $dist(P_i, P_l)$ is the space interval between P_i and P_l.

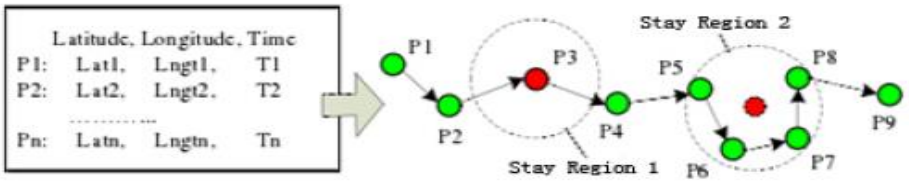


Figure 1. GPS Log and Stay Regions

 Copyright ⓒ 2015 SERSC

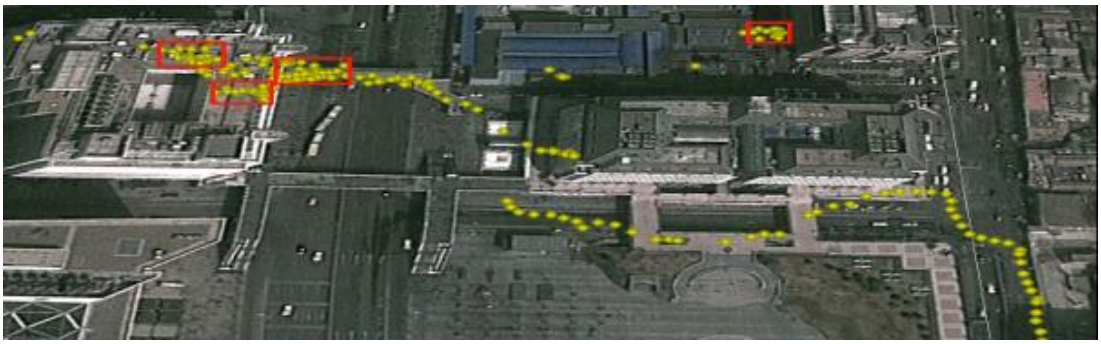


Figure 2. Stay Regions

A stay region stands for a geographic region where a user stays for a while, e.g., the restaurants and shopping malls, etc [11]. Two types of stay region as depicted in Fig.1. *Stay region* 1, the user maintains stationary at P_3 for over a time threshold. *Stay region 2*, the user wanders around within a spatial region for over a time threshold. [12] Stay region is a point set that Euclidean distance and time interval between any two points within a certain realm. In this paper，the representation is a rectangular whose lower left corner as point (*P_l*) and upper right corner as point (*P_r*), as shown in Fig.2.

Definition 3 (Time interval). *Time_ interval* is the temporal span of the stay region. The minimum and maximum times of the region are the entering time (*T_arr*) and the leaving time (*T_lea*).

2.2. Generate STIs from Stay Regions

Every element of the STI as mentioned above, the format of the STI can be expressed as Table 1 and the Table 2 is an example of STI.

Table 1. Format of STI

Id	User_id	Stay_ region		Time_ interval	
id	*u_id*	*P_l*	*P_r*	*T_arr*	*T_lea*

Table 2. STI Generating

Id	User_id	Stay_ region		Time_ interval	
1	003	39°59'28.66", 116°19'37.43"	39°59'28.70", 116°19'37.56"	*8：00*	*8：07*

3. STIs Mapping

In this subsection, we will discuss how to map the longitude, latitude and time of the STI to the 3D-grids. Inspired by 2D-grids, temporal information is integrated into 2D-grids and the 3D-grids are proposed. Mining FSTIs is to find out these adjacent 3D-grid cells which meet the minimum support threshold σ. Then the adjacent cells are combined to recombine the FSTIs.

3.1. Set the 3D-grid Cells

In fact, every 3D-grid cell has *x*-axis, *y*-axis and z-axis which correspond with longitude, latitude and time. Since FSTIs are expected to repeat themselves hourly, daily and monthly, etc, the temporal dimension of the FSTIs is projected down to the minutes-of-hour, hours-of-day or days-of-month, ect. The periodicity can be facilitated by projections of the temporal domain to appropriate finer or coarser levels of granularity.

Copyright ⓒ 2015 SERSC

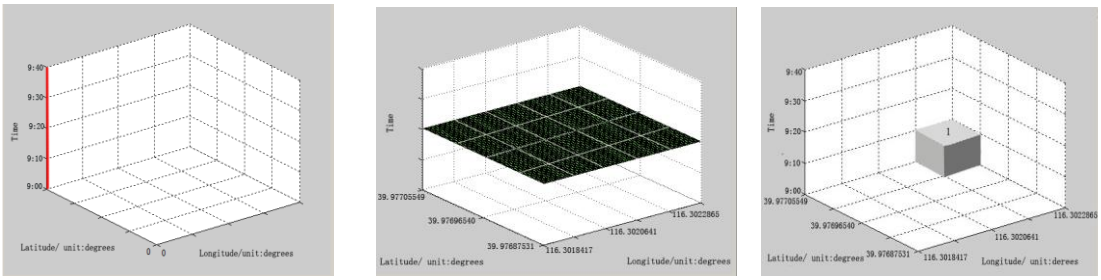
[7] The z-axis length of the cell is set according to periodicity and it is set arbitrarily less than one cycle. For example, the periodicity is "*day*", so the z-axis length of the cell can be set in less than one day arbitrarily, as shown in Figure 3 (a) as "10 *minutes*". At the same time, the range of the frequent spatial region should be logical. So, the x-axis and y-axis lengths of the cell should be set referring to geography, as shown in Figure 3 (b).

3.2. Map STIs to 3D-grids

Obviously, all of the cells which the STIs covered are mapped. For example, we can map the STI in Table 3 to the 3D-grids whose x-axis and y-axis lengths of the cell are 5 *meters* and z-axis length of the cell is 10 *minutes*. As shown in Figure 3 (c), this STI just cover one cell. In Fig. 4 the Map algorithm is given.

Table 3. A STI

Id	*User_id*	*Stay_ region*		*Time_ interval*	
6	003	39.97704374, 39.97700012	116.3022222, 116.3021003	9：02	9：03



(a) Temporal dimension **(b) Spatial dimension** **(c) Covered cell**

Figure 3. The 3D-grid Cells

Algorithm 1: Map（*Trajectory, cell*）

Input: dataset of trajectory(*Trajectory*), *cell*
Output: cell mapped times
1. Scan(*Trajectory*)⟶∑*Stay_ regions*={P_1, P_2,...,P_m}**/*Step 1. Generate STIs*/**
2. ∑*Stay_ regions*={ P_1,P_2,...,P_m }⟶∑*STIs*
3. for each STI do **/*Step 2. Map to 3D-grids and record Count(M）*/**
4. Get(*cell.start*) and Get(*cell.number*)
5. *M*=0; for(*i=cell.start; i<= cell.start+cell.number; i++*)
6. {*M=M*+1}
7. end for
8. end for

Figure 4. Map Algorithm

4. Extraction-merger Strategy

Definition 4 (FSTI). A transaction set *Transaction*= {T_1, T_2, …, T_n}, item sets T_i={I_1, I_2, …, I_n}, the support of an item is defined as the percentage of transactions that contain the item in the transaction set and it is written as *support* (I_i). An item is frequent if its support is not less than the user-specified minimum support threshold σ.

In fact, to mine FSTIs is to find the STIs whose supports are not less than the minimum support threshold σ. We have map the STIs to the 3D-grids described in Section 3 and the parameter *M* is set to record the mapped time of the each grid cell. Because the STIs are

 Copyright ⓒ 2015 SERSC

known initially and the periodicity is set by user, the number of the transactions *Count(T)* can be known easily. So, the support of the each grid cell can be calculated as Formula (1). The next step is to find out these adjacent 3D-grid cells which meet the minimum support threshold σ and then combine the adjacent 3D-grid cells to recombine the FSTIs. The Extraction - merger algorithm is given in Fig. 5.

$$support\ (cell) = \frac{M}{Count(T)} \tag{1}$$

Algorithm 2: Extraction - merger（σ, *cells*）

Input: support threshold σ, *cells*
Output: FSTIs
1. for each *cell* do/***Step 1. Find out all core cells*/**
2. if *support* (*cell*) ≥ σ then
3. label *cell* as core cell
4. end for
5. *Rcore*=*Φ*/***Step 2. Find out all core regions*/**
6. for each *cell* do
7. if $cell_i$ is a core cell then
8. r = new core-region({c})
9. while true do
10. if ∃ a core *cell* $c' \in r$.neighbours then
11. $r = r \cup \{c'\}$
12. else
13.Rcore=Rcore∪{r}
14. end for
15. Rcore⟶FSTIs **/*Step 3. Representation of Core regions */**

Figure 5. Extraction - Merger Algorithm

5. Experiment

5.1. Select Dataset

In this paper, the GPS trajectory data is collected in (Microsoft Research Asia) Geolife project by 178 users in a period of over three years. A GPS trajectory of this dataset is represented by a sequence of time-stamped points, each of which contains the information of latitude, longitude and altitude. This dataset contains 17,621 trajectories with a total distance of about 1.2 million kilometers and a total duration of 48,000+ hours. [11, 13-14]. The GPS trajectory data of one user is chosen as subject in the experiment.

5.2. Set the Parameters

This paper is based on a density-based clustering that is self-adaptive recognition method of the stay area [10]. The complexities of time and space in that method are relatively moderate. In that paper, firstly, trajectory is divided into trajectory segments according to different modes of transportation based on a change point-based segmentation method. Then, walk segment and non-walk are distinguished based on Back Propagation neural network. Finally, cluster the foot segments. So the *Stay_ region*s can be recognized and the STIs can be generated. According to people's habits, we select the time granularity as "*minutes of day*".

As is known to all, the unit length of the longitude is decreasing as the latitude escalating. We can calculate the unit length of the longitude at different latitudes as Formula (2). *B* is the latitude of the point. *R* is the equatorial radius *R*=6378137*meters* and *r* is the radius of polar *r*=6356752*meters*. Formula (3) can be used to calculate the longitude interval *N* that corresponding fixed length *D*. The unit length of the latitude is

Copyright ⓒ 2015 SERSC

about 111000*meters* at any longitude. The Formula (4) is used to calculate the latitudes interval *M* that corresponding fixed length *D*.

$$S = \frac{\cos B \times (R - B \times (R - r)/90) \times 2\pi}{360} \quad (2)$$

$$N = D/S \quad (3)$$

$$M = D/111\,000 \quad (4)$$

5.3. Result and Analysis

The result of the experiment is that the trajectories were mapped to 103992 grid cells. Some FSTIs with different supports are shown in Table 4. When the movement of a user is more regularly, the scale of STIs is smaller and the supports of the FSTIs are higher. Based on this method, we can mine the FSTIs with different supports. Among them, some examples of FSTI in trajectory data discovered by real dataset are shown in Fig.6.

Table 4. Some FSTIs with different Supports

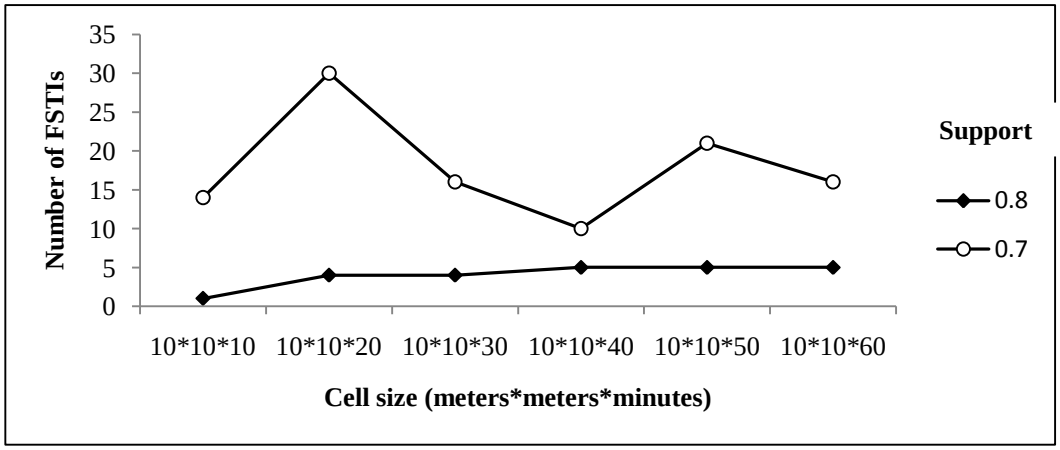
Id	*User_id*	*Stay_ region*		*Time_ interval*		*Count (T)*	*M*	*support*
1	003	39.9993978, 116.322192	40.0011996, 116.3224144	9:50	10:00	29	24	0.827586
2	003	39.9993978, 116.3220808	40.0011996, 116.3271962	9:50	10:10	29	23	0.793103
3	003	39.9993978, 116.3218584	40.0057041, 116.3259729	9:50	10:10	29	22	0.758621

Figure 6. Some Examples of FSTI in Trajectory Data Discovered

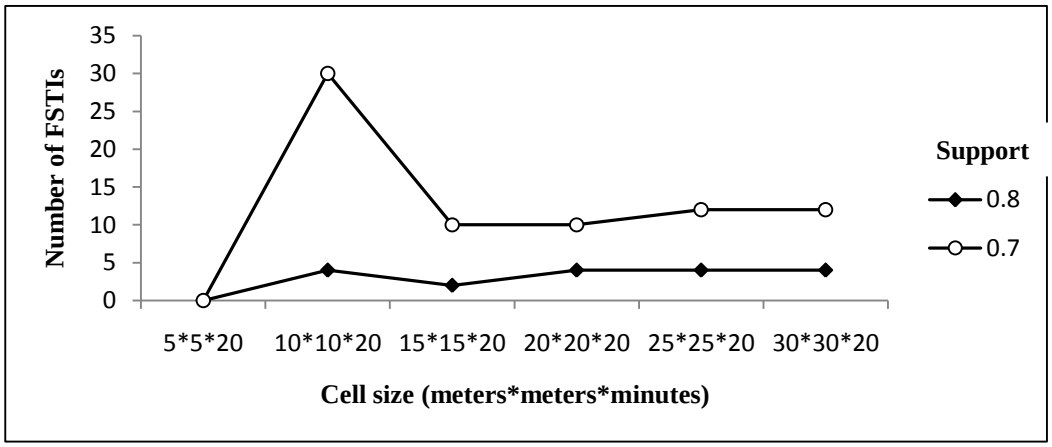
The gird cell size has influence on the mining result. Fig.7 summarizes the number of FSTIs with different supports versus cell size. As shown in Fig.7 (a), the number of FSTIs gets a high as 10meters*10meters*20minutes and gets a low as 10meters*10meters *40minutes. The reason is that when the z-axis length of the cell is set accord with temporal regularities of the user, the FSTIs will be mined easier like 10meters*10meters*20minutes. Moreover, when some approximate trajectory locations are close to the boundary of predetermined grid cells, they will be most likely assigned to different cells by the strict boundary constraints. Therefore, some potential meaningful FSTIs will not be discovered due to this issue like 10meters*10meters*40minutes. As shown in Fig.7 (b), the number of FSTIs with support 0.7 decreases from its peak of 10meters*10meters*20minutes. The reason is that when the x-axis and y-axis lengths of the cell are set accord with the size of the stay region, the FSTIs will be mined easier.

 Copyright ⓒ 2015 SERSC

Moreover, when the cell size is set too small, every cell will be mapped less and its support will decrease. So the FSTIs will be undiscovered like 5meters*5meters*20minutes. After 15meters*15meters*20minutes, the size of stay region increases, the number of FSTIs will remains little change, but the precision of the stay region will decrease. The number of FSTIs with higher support is insensitive to the cell size and 10meters*10meters*20minutes is the most appropriate grid cell size in our experiments.



(a) Different Temporal Dimensions



(b) Different Spatial Dimensions

Figure 7. Influence of Cell Size on the Mining Result

6. Conclusion and Future Work

This work presents novel definitions of STI and FSTI integrated spatial and temporal attributes to the time aspect ignored in ROIs or hot regions. An approach for mining FSTIs is developed correspondingly. The FSTIs can represent a moving object often visits which area in what time, which can provide more useful information to improve the level of the location-based services(LBS). 3D-grids are employed to map the STIs and the extraction - merger strategy is used on the frequent grid cells to recombine the FSTIs. Experimental results on a large number of actual trajectory data verify the effectiveness and feasibility of the proposed algorithm.

In the future work, it is extensible to explore appropriate cell size because it has an effect on the mining results and the sharp boundary problem needs to be solved. Moreover, a possible combination can integrate with space geographic knowledge and road networks,

Copyright ⓒ 2015 SERSC

in order to obtain a more intuitive space region with space semantics. Additionally, some frequent spatio-temporal patterns and Spatial-Temporal Association Rules (STARs) which are based on the FSTIs can be mined to find regularities of a user or a group of users.

References

[1] K. E. Liu, J. C. Xiao, Z. M. Ding and M. S. Li, "Found the Hot Areas in Database of Trajectory", Journal of software, vol.8, **(2013)** , pp.1816-1835.
[2] Y. Liu and C. W. Sun, "An analysis method of multi-granularity hotspots based on the clustering of stay point", Microcomputer information, vol.9, **(2012)**, pp.295-297.
[3] R. Agrawal, T. Imielinski and A. Swami, "Mining association rules between sets of items in large databases", Proceedings of the ACM SIGMOD International Conference on Management of Data, **(1993)**.
[4] J. Han, J. Pei and Y. Yin, "Mining frequent patterns without candidate generation", Proceedings of the 2000 ACM-SIGMOD international conference on management of data (SIGMOD'00), **(2000)**; Dallas, TX.
[5] L. Wang, K. Y. Hu, T. Ku and X. H. Yan, "Mining frequent trajectory pattern based on vague space partition", Knowledge-Based Systems, vol.50, **(2013)**, pp.100-111.
[6] L. Chen, M. Q. Lv, Q. Ye, G. C. Chen and J. Woodward, "A personal route prediction system based on trajectory data mining", Information Sciences, vol.181, no.17, **(2011)**, pp.1264-1284.
[7] G. Gidófalvi and T. B. Pedersen, "Mining Long, Sharable Patterns in Trajectories of Moving Objects", GeoInformatica, vol.13, no.1, **(2009)**, pp.27-55.
[8] H. P. Cao, N. Mamoulis and D. W. Cheung, "Mining Frequent Spatio-temporal Sequential Patterns", Proceeding of: Data Mining, Fifth IEEE International Conference on Data Mining (ICDM'05), icdm, **(2005)**.
[9] G. Gidófalvi and T. B. Pedersen, "Spatio–temporal Rule Mining: Issues and Techniques", Proceedings of the 7th International Conference on Data Warehousing and Knowledge Discovery, **(2005)**; DaWaK, LNCS.
[10] P. Y. Bi, L. Shen and X. Li, "Self-Adaptive Recognition Method of the Stay Area in the Passenger Flow Trajectory", ICIC Express Letters, Part B: Applications, vol.3, **(2014)**, pp. 891-896.
[11] Y. Zheng, L. Z. Zhang, X. Xie and W. Y. Ma, "Mining interesting locations and travel sequences from GPS trajectories", Proceedings of International conference on World Wild Web, **(2009)**; Madrid Spain.
[12] Y. Ye, Y. Zheng, Y. K. Chen, J. H. Feng and X. Xie, "Mining Individual Life Pattern Based on Location History", In proceedings of the International Conference on Mobile Data Management, **(2009)**.
[13] Y. Zheng, X. Xie and W. Y. Ma, "GeoLife: A Collaborative Social Networking Service among User, location and trajectory", Invited paper, *in* IEEE Data Engineering Bulletin, vol.33, no.2, **(2010)**, pp.32-40.
[14] Y. Zheng, Q. N. Li, Y. K. Chen, X. Xie and W. Y. Ma, "Understanding Mobility Based on GPS Data", Proceedings of ACM conference on Ubiquitous Computing, **(2008)**; Seoul, Korea.

 Copyright ⓒ 2015 SERSC