Periodic Pattern Mining Based on GPS Trajectories

Xiaopeng Chen[1,a], Dianxi Shi [2,b] , Banghui Zhao[3,c] and Fan Liu[4,d]

[1,2,3,4]National Laboratory for Parallel and Distributed Processing, School of Computer,

National University of Defense Technology, Changsha 410073, China

[a]manco1314@aliyun.com, [b]dxshi@nudt.edu.cn, [c]zhao37890@126.com, [d]liufan14@nedt.edu.cn

Keywords: Mobile Data Mining; GPS Trajectories; Periods Detection

Abstract. With the rise of LBS (Location Based Service), lots of user recommendation system based on location trajectory analysis has emerged. Finding periodic behavior is essential to analysis user's activity. In order to solve the problems with trajectory outliers and human intervention in periodic parameters, we propose a three-stage framework called PPM (Periodic Pattern Mining) to detect periodic pattern based on people's trajectory. First of all, we preprocess the trajectory data to extract stay points. Secondly, the sequence of stay points are clustered to construct the important places. At the last stage, the movement sequence is transformed into a binary sequence. Then period of every binary sequence is detected by a probabilistic model. The experiment based on public mobile dataset shows that the proposed method can be used to mine people's periodic activity pattern effectively.

Introduction

Recently, smart phones and wearable devices spread rapidly all around the world. The rise of these mobile terminals make it easier for people to obtain their positions enabling numerous trajectories collected. As these trajectories contain an abundant amount of knowledge, trajectory data mining has drawn great attention of plentiful experts and scholars, and it has become one of the current research hotspot. The geographical location information and trajectory data collected has a wide range of applications greatly improving users' experience, like path navigation, nearby food recommendation etc. Based on the GPS data, we can also dig out the transportation modes [1]. Transportation mode is an important attribute to describe the behavior of people, usually including walking, bicycle, bus, car etc. The results mined from the GPS data can provide reasonable route for people. In these applications, lots of low level applications are converted to stay location, and high level of schema information is used in the analysis of mobile phone users. Thus, not only the daily activities pattern but also high level potential information can be mined from GPS trajectories.

As one of the necessary equipment carried by modern people every day, mobile phone records the user's daily trajectory, so we can find out people's daily movement pattern from the GPS trajectories. With the fact that periodicity is a phenomenon which occurs frequently in our daily life, periodic activities pattern mining can be used to further understand our daily habits and customs. However, in the field of periodic activity pattern mining exists the problems of uncertainty data sampling frequency and data sparseness. And in order to solve the problems above, we propose a method to detect the periodic activity pattern which is based on probabilistic model. In this paper, we try to make a research on mining people's periodic activity pattern from their historical trajectory.

The rest of the paper is organized as follows. We discuss related work in Section 2. Section 3 outlines the general framework. In Section 4, we introduce the data preprocessing procedure. Section 5 describes the method to discover the activities period. The experimental discussions are provided in Section 6. Summary is in Section 7.

Related Work

Data mining [2] refers to the process that extracting the implicit information and knowledge from the large, incomplete, fuzzy, random and practical data which people do not know in advance, but is potentially useful. Mobile data mining [3] is the discovery of useful knowledge from mobile phones. As one of the pioneers in the study of user mobility patterns, in 1995, Liu et al. [4] propose a simple

linear model to model the user's mobile model. After that, more and more new models are put forward, for example, Liang et al. [5] use the Gauss Markov model, to improve the accuracy of the model. Wang et al. [6] mine time-space similar groups according to the records of user locations. In the foundation of location study, Yava et al. [7] mine the sequence pattern based on the user's mobile trajectory, and predicts the area that the user is about to enter. Hwang et al. [8] propose a method to get the group of similar users by the movement pattern mined based on the user's trajectory. Tseng et al. [9] propose a mining algorithm for instantaneous moving sequence patterns, which is based on the user's mobile path and time interval. Fang et al. [10] propose a binary mining algorithm mining the association rules of the spatial location based on spatial database. Based on the prior efforts, mining user's behavior patterns using trajectory data is becoming an important research problem. Lots of previous work have been done to discover and mine users' movement patterns [11,12]. Zheng et al. [13] propose a method to get the interesting tour place and the classic visit sequence in specific areas by mining the GPS trajectory of multiple users. Monreale et al. [14] mine out the user's universal movement pattern of an area, and predict the user's next location. Zhu et al. [15] use instantaneous entropy to define the user's moving speed, and mine the user's frequent location, frequent path, meaningful location and movement pattern, etc. Chen et al. [16] select four aspects data of user information, mobile location, residence time, and requested services, mining the user's movement patterns and designing relevant prediction.

Compared with traditional methods on movement pattern mining, our method is based on detected movement period with following advantages. Firstly, it can detect period automatically according to GPS data instead of human input. Secondly, a probabilistic model is adopted to detect periods. Our proposed method has thoroughly considered the uncertainties and noises in periodic activities.

PPM Framework Overview

In order to availably find periodic activity, we design and implement PPM (Periodic Pattern Mining) framework as Fig. 1 which consists of mobile terminal and server side.

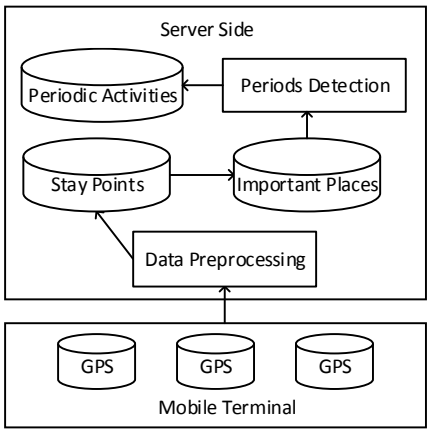

Fig. 1. PPM Framework.

The mobile terminal is used to collect GPS trajectory data, then upload data to server side for the following mining procedure. On the other hand, server side is our main component of PPM framework aiming at mine periodic activity based on the collected GPS data.

The raw trajectory data is denoted as $P=\{p_i|0<i\leqslant n\}$, in which p_i is $p_i(Lat,Lg,T)$. Parameters are latitude, longitude, and time information. Given GPS trajectory, our problem aims at mining the user's important places and the corresponding activity period. Before giving the meaning of "important place", we firstly introduce another term, *stay point*.

DEFINITION 1 (STAY POINT SEQUENCE S). A stay point is a geographical area where a user wandered for a period of time including places that the user stayed static over a span. It is denoted as $S=\{s_1,s_2,\ldots,s_m\}$, in which $s=s(Lat,Lg,T_a,T_l)$, T_a and T_l is the arriving and leaving time in 's' point. For instance, points that lie within a physical region where a user maintained stationary state more than a

time threshold like a building or where the user wanders over a distance threshold like a campus are all stay points.

However, there is not one-to-one match between each stay point and a geographical location with semantic information. The same single place do not have the changeless GPS because the collected GPS is often different in different time. Speaking simply, no two point in trajectory data is exactly the same. So we need to further cluster points within the same location as the important place.

DEFINITION 2 (IMPORTANT PLACE SEQUENCE I). As stay point do not have explicit semantic meaning, it is necessary to cluster stay points into important place sequence. Each important place is a geographical location with semantic information like company etc. as Fig. 2 describes.

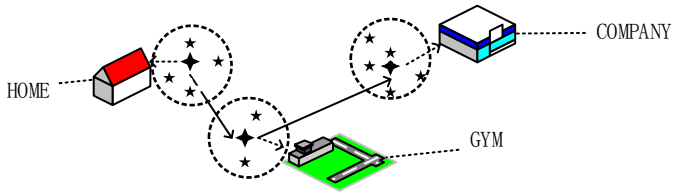

Fig. 2. Sketch Map of Important Place

Preprocessing Data

In this section, we preprocess the trajectory data to get important places for the period detection. This includes two subtasks, namely, extracting stay points and finding important place.

Filtering Noise. The collected GPS data usually contains outliers, due to sensor noise and other factors like poor signals in some areas. So it is important to filter noise data in the trajectories. We use velocity filter and accelerator filter to clean up the outliers before extracting stay points.

We calculate every two consecutive GPS points' velocity value to remove the points outside the velocity threshold and calculate every three consecutive GPS points' accelerator value to remove the points outside the accelerator threshold. According to the restriction of people movement mode, we set the speed threshold as 150km/h and the accelerator threshold as $15m/s^2$.

Extracting Stay Points. According to the above notion, the stay point is determined by time threshold T_{thresh} and spatial distance threshold D_{thresh}. We focus on physical region where a user hang around for over a certain time interval, whose central point is regarded as a stay point. Supposing $P=\{P_i,P_{i+1},\ldots,P_j\}$ is a series of points contained in a stay point s, for $\forall x$, where $i<x\leqslant j$, Distance(p_m,p_i) $\leqslant D_{thresh}$, and $p_j.T-p_i.T\geqslant T_{thresh}$, a stay point $s(Lat,Lg,T_a,T_l)$ can be calculated by the following formula.

$$s.Lat=\sum_{x=i}^{j}p_x\times Lat/|P|, s.Lg=\sum_{x=i}^{j}p_x\times Lg/|P|, s.T_a=p_i.T, s.T_l=p_j.T \quad (1)$$

Constructing Important Places. In this thesis, we use density-based clustering algorithm to extract important places. Unlike other clustering algorithm, such as K-means algorithm which needs to specify the number of cluster centers, or grid-based partition method who may divide GPS points of the same place in different grids causing boundary problems, density-based method can find clusters with irregular structure. OPTICS is a density-based algorithm who improves the famous DBSCAN algorithm. We use a method similar to OPTICS to cluster important places.

We extract an unprocessed stay point and iterate all current cluster set to test if it belongs to a cluster, if it does, then update the cluster, and if it does not, we construct a new cluster set with it. At last, we remove cluster whose points number less than the given count threshold.

Detecting Period

Transforming Sequence. Next, we detect period for every important place. Take an mined important place I_m as an example, we transform the moving sequence to a binary sequence $B=b_1b_2\ldots b_n$, where $b_i=1$ when the user is within the important place at timestamp i and $b_i=0$ otherwise.

Periodic Probabilistic Model. In this thesis, we try to provide a probabilistic model to detect the period of user trajectories, which can solve problems like acquisition frequency uncertainty.

DEFINITION 3 (PERIODIC BINARY SEQUENCE). Supposing that X=x(t) (0<t<n) is a sequence, if $\exists T \in Z$, satisfy x(t+T)=x(t), then we call X a periodic sequence. Furthermore, if the vector $p=(p_1,p_2,\ldots,p_T)$, ($p_i \in [0,1]$) is a periodic distribution vector with length T, and X is independent and subject to Bernoulli distribution with p, then X is a periodic binary sequence satisfying the periodic distribution vector p.

For the above binary sequence $B=b_1b_2\ldots b_n$ transformed from important places, we aim to find period T who satisfy B as a periodic binary sequence.

DEFINITION 4. Supposing that X=x(t) (0<t<n) is a periodic binary sequence, X^+ represents the set of independent variable t satisfying x(t)=1, and X^- represents the set of independent variable t satisfying x(t)=0. We suppose that I_t is the power set of [0:T-1], and T is a period candidate. To $\forall I \in I_t$, we give the equation as followed:

$$X_I^+ = \{t \in X^+ : F_T(t) \in I\}, X_I^- = \{t \in X^- : F_T(t) \in I\},\ F_T(t) = \mathrm{mod}(t,T) \tag{2}$$

THEOREM 1. To a periodic binary sequence X=x(t) with length n who is subject to periodic distribution vector p, we have the following equation:

$$\lim_{n\to\infty} \beta_x^+(I,T) = \frac{\sum_{i\in I} p_i}{\sum_{i=0}^{T-1} p_i}, \lim_{n\to\infty} \beta_x^-(I,T) = \frac{\sum_{i\in I}(1-p_i)}{\sum_{i=0}^{T-1}(1-p_i)}, \beta_x^+(I,T) = \frac{\left|X_I^+\right|}{\left|X^+\right|}, \beta_x^-(I,T) = \frac{\left|X_I^-\right|}{\left|X^-\right|} \tag{3}$$

Equation (7) can be demonstrated using Bernoulli distribution and law of large numbers. Next, we detect the moving period based on theorem 1. $\forall I \in I_t$,

$$\Delta x(I,T) = \beta_x^+(I,T) - \beta_x^-(I,T) \tag{4}$$

If T is the detected period, then we have the equation:

$$\alpha(t) = \max_{I\in I_T} \Delta(I,T), 0 \le \alpha_x(t) \le 1 \tag{5}$$

If T totally satisfies the time sequence with period T, then α (T)=1.

Detecting Periods. Given the binary sequence converted from the important place sequence, we treat it as a sequence who satisfied some periodic distribution vector. According to the description in above section, we compute every periodic candidate T for every binary sequence. Of all the periodic candidate T, we choose the one who has the biggest probability as the binary sequence period, which is the period of the trajectory. We will first describe the detecting method as Fig. 3.

Algorithm Period Detect, PD(X, T_{max})

Input: 0-1 sequence X, time threshold T_{max}
Output: Period T
1. T_0=1, map; //new a map container
2. while $T_0 < T_{max}$ do,
3. $P = \sum_{i\in I^*} c_i, I^* = \{i \in [0, T_0 - 1] : c_i > 0\}$
4. $c_i = \frac{p_i}{\sum_{k=0}^{T-1} p_k^T} - \frac{q_i}{\sum_{k=0}^{T-1} q_k^T}$
8. map.add(T_0,P);
9. if T_{can} is max candidate T in map then //get max T in map->getMaxT
10. T=T_{can};
14. return T;

Fig. 3. Alrorithm of Period Detection

Experiment

There are several public mobile context datasets like Device Analyzer Dataset [17] and Geolife 1.2 [18]. In our research, we use Geolife 1.2 as our experimental data. The running environment of this experiment is Windows 7 operating system with 4GB memory, and the algorithm is written using Java language. The map displays using the Javascript of Baidu Map API.

Clustering Important Places. We select NO.151 experimenter (Tom) as an example to mine his period patterns. Tom has 419344 GPS logs in 2011. Fig. 4 shows 364503 GPS points after filtering. Fig. 4 (a) is the global map of trajectories, Fig. 4 (b) is the local map of trajectories in Beijing. We can see that Tom has travelled several areas in 2011 as Beijing, Shanxi Province, Chengdu and Hangzhou.

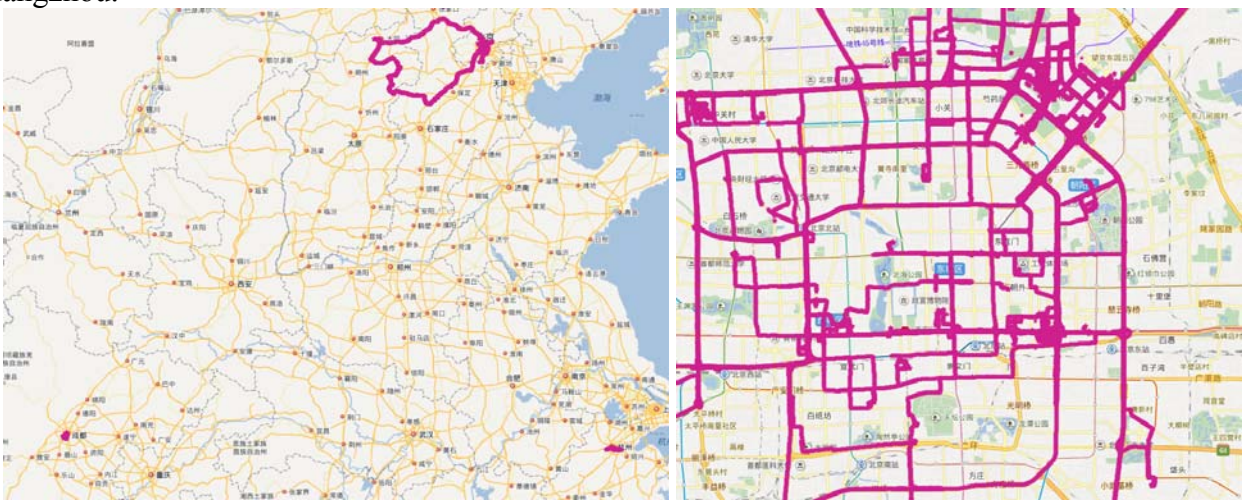

Fig. 4. (a) Global Trajectory Map after Filtering, (b) Local Trajectory Map after Filtering

Fig. 5 shows Tom's stay points mined using SPD (Stay Points Detecting) algorithm assuming that the time threshold is 15min and the distance threshold is 200m. After detecting, we get 345 stay points. Fig. 5 (a) is the global map of trajectories, Fig. 5 (b) is the local map of trajectories in Beijing.

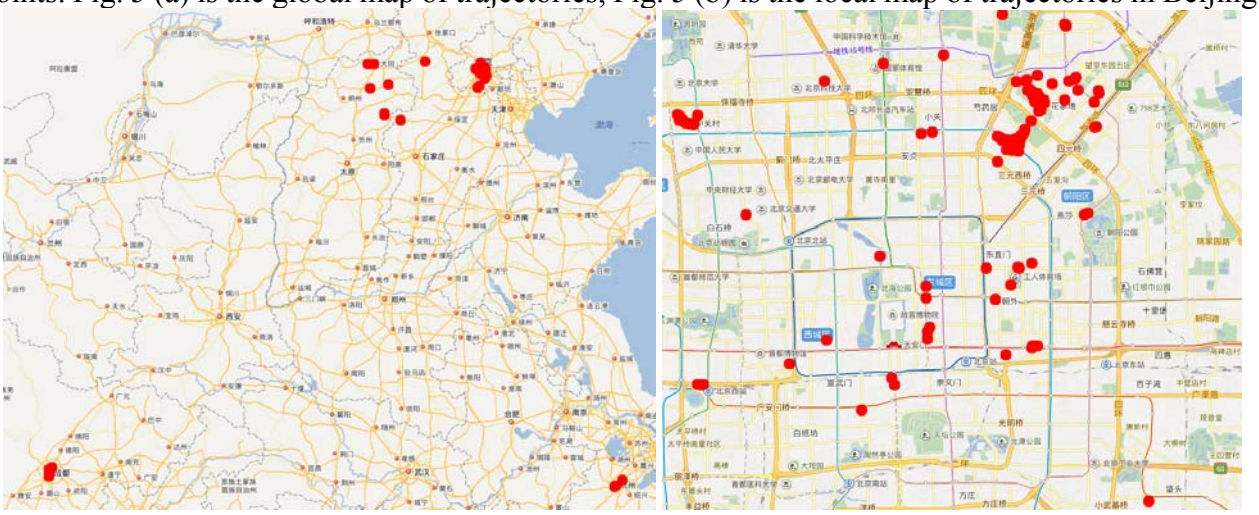

Fig. 5. (a) Global Map of Extracted Stay Points, (b) Local Map of Extracted Stay Points

After further processing to cluster stay points, we get the clustered important places. Table 1 shows the results of important places.

Table 1 The Clustered Important Places

Reference spot ID	Longitude	Latitude	Number of stay points
1	116.4153416	39.9818752	252
2	104.0728018	30.5903695	8
3	120.1869724	30.2799014	4
4	113.6026764	39.0124993	3
5	113.1445734	40.1157183	5
6	116.4419583	39.9394337	10
7	116.3983592	39.9430833	11
8	116.4822236	39.9561279	2

From Table 1, we can see that clustering 1 contains the most stay points, meaning that clustering 1 is the place Tom frequently visits. It is not difficult to find clustering 1 is Beijing from the map.

Detecting Periods. First of all, we get the binary movement sequence by dividing Tom's 2011 trajectory into 17520 sections in every 30 minutes. If the section contains a stay point, the value of the section is 1, and 0, otherwise. By using the period detection algorithm, we detect the period of every binary sequence. The periods detected for each important place are shown in Table 2.

Table 2 Detected Periods

Reference spot ID	Number of stay points in the clustering area	Period length (unit: hour)
1	252	24
2	8	716
3	4	722
4	3	167
5	5	168
6	10	168
7	11	168
8	2	169

From Table 2, we can see that important place 1 has period of 24 hours, So place 1 has the period which can be regarded as one day. In other words, Tom has daily movement pattern in place 1. Similarly, Tom has weekly movement pattern in place 4,5,6,7,8 and monthly movement pattern in place 2,3. Furthermore, we can find that place 1 is in Beijing City, place 2 in Chengdu City, place 3 in Hangzhou City, place 4,5,6,7,8 in Shanxi Province. On the other hand, one daily, one weekly and one monthly behavior pattern conform to people's living habits. It can be speculated that Tom is likely to be a company staff who often travels between Beijing and Shanxi Province, and he goes to Chengdu and Hangzhou on business for several times in 2011.

Summary

In this paper, we propose a three-stage framework to detect periodic pattern based on GPS trajectory. In the data preprocessing procedure, we propose the velocity filter and accelerator filter to clean up the outliers solving the problem of trajectory outliers. Furthermore, we give a method to extract stay points and then use clustering algorithm to construct important places, with the reason that different important places can overlap in time, so we solve another problem of multi-period cross-cutting. In addition, we use a probabilistic model to detect periods, and solve the problem of data sampling frequency uncertainty. We also solve the problem of human intervention in periodic parameters in mining human mobility patterns. The method is provably robust to incomplete observations and sparse data.

This thesis lies down a solid foundation for future works towards individual movement knowledge mining from trajectories. And we could also to further improve our algorithm as the time complexity of PPM is O (n*n) in the worst case.

Acknowledgements

We are thankful to Microsoft Research Asia for providing us Geolife dataset. The first author would like to thanks Guangfen Zhu for her valuable comments on this work.

References

[1] Yu Zheng, Yukun Chen, Quannan Li. Understanding transportation modes based on GPS data for Web applications. ACM Transaction on the Web. 4(1), January, 2010. 1-36.

[2] Guiqin Wang, Dao Huang. The Summary of The Data Mining Technology[J].2007.

[3] Goh J Y, Taniar D. Mobile data mining by location dependencies[M] Intelligent Data Engineering and Automated Learning–IDEAL 2004. Springer Berlin Heidelberg, 2004: 225-231.

[4] Liu G Y, Maguire Jr G Q. A predictive mobility management algorithm for wireless mobile computing and communications[C] Universal Personal Communications. 1995. Record., 1995 Fourth IEEE International Conference on. IEEE, 1995: 268-272.

[5] Liang B, Haas Z J. Predictive distance-based mobility management for PCS networks[C] INFOCOM'99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE. IEEE, 1999, 3: 1377-1384.

[6] Wang Y, Lim E P, Hwang S Y. On mining group patterns of mobile users[C] Database and Expert Systems Applications. Springer Berlin Heidelberg, 2003: 287-296.

[7] Yavaş G, Katsaros D, Ulusoy Ö, et al. A data mining approach for location prediction in mobile environments[J]. Data & Knowledge Engineering, 2005, 54(2): 121-146.

[8] Hwang S Y, Liu Y H, Chiu J K, et al. Mining mobile group patterns: A trajectory-based approach [M] Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2005: 713-718.

[9] Tseng V S, Lu E H C, Huang C H. Mining temporal mobile sequential patterns in location-based service environments[C] Parallel and Distributed Systems, 2007 International Conference on. IEEE, 2007, 2: 1-8.

[10] Fang G, Wei Z K, Yin Q. Extraction of spatial association rules based on binary mining algorithm in mobile computing[C] Information and Automation, 2008. ICIA 2008. International Conference on. IEEE, 2008: 1571-1575.

[11] Tseng V S, Lu E H C, Huang C H. Mining temporal mobile sequential patterns in location-based service environments[C] Parallel and Distributed Systems, 2007 International Conference on. IEEE, 2007, 2: 1-8.

[12] Chen T S, Chou Y S, Chen T C. Mining user movement behavior patterns in a mobile service environment [J]. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 2012, 42(1): 87-101.

[13] Zheng Y, Zhang L, Xie X, et al. Mining interesting locations and travel sequences from GPS trajectories[C] Proceedings of the 18th international conference on World wideweb. ACM, 2009: 791-800.

[14] Monreale A, Pinelli F, Trasarti R, et al. Wherenext: a location predictor on trajectory pattern mining[C] Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009: 637-646.

[15] Zhu Y, Zhang Y, Shang W, et al. Trajectory enabled service support platform for mobile users' behavior pattern mining[C] Mobile and Ubiquitous Systems: Networking & Services, MobiQuitous, 2009. MobiQuitous' 09. 6th Annual International. IEEE, 2009: 1-10.

[16] Chen T S, Chou Y S, Chen T C. Mining user movement behavior patterns in a mobile service environment [J]. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 2012, 42(1): 87-101.

[17] Wagner D T, Rice A, Beresford A R. Device Analyzer: Large-scale mobile data collection[J]. ACM SIGMETRICS Performance Evaluation Review, 2014, 41(4): 53-56.

[18] Y. Ye, Y. Zheng, Y. Chen, et al, Mining individual life pattern based on location history. the 10th International Conference on Mobile Data anagement, Taipei, Taiwan, 2009: 1-10P

