A novel route restoring method upon geo-tagged photos

Guannan Wang[12], Zhizhong Wang[1], Zhenmin Zhu[2], Saiping Wen[2]
[1] School of Mathematics and Statistics, Central South University
HuNan, China,
[e-mail: wgn1103@gmail.com]
[2] Pervasive Computing, Institute of Computing Technology, ChineseAcademy of Sciences, Beijing Key Laboratory of Mobile Computing andPervasive Device
No.6 Ke Xue Yuan South Road,Haidian District, Beijing, China
[e-mail: zmzhu@ict.ac.cn]
*Corresponding author: Zhenmin Zhu

Received August 13, 2012; revised February 16, 2013; accepted May 9, 2013; published May 31, 2013

Abstract

Sharing geo-tagged photos has been a hot social activity in the daily life because these photos not only contain geo information but also indicate people's hobbies, intention and mobility patterns. However, the present raw geo-tagged photo routes cannot provide information as enough as complete GPS trajectories due to the defects hidden in them. This paper mainly aims at analyzing the large amounts of geo-tagged photos and proposing a novel travel route restoring method. In our approach we first propose an Interest Measure Ratio to rank the hot spots based on density-based spatial clustering arithmetic. Then we apply the Hidden Semi-Markov model and Mean Value method to demonstrate migration discipline in the hot spots and restore the significant region sequence into complete GPS trajectory. At the end of the paper, a novel experiment method is designed to demonstrate that the approach is feasible in restoring route, and there is a good performance.

Keywords: Geo-tagged Photo; Hot spot; Travel route; Route restoring; Trajectory mining;

http://dx.doi.org/10.3837/tiis.2013.05.017

1. Introduction

Recent advancements of information and location-aware technologies have enhanced our capability of collecting individual trajectory data of people, vehicles, or other moving objects. On one hand, a branch of geographic applications based on user-generated trajectory data has appeared on the Web and received considerable attention, such as Yahoo travel [1] and GPS Track log route exchange Forum [3]. These webs also provide travel services including hot spots, trajectory, and hotel recommendation, there are also some applications to do something useful to the traffic ([5], [6]). However, the online travel services in the nowadays pay more attention on the reservation system for accommodation but not the Location Based Service (LBS), not to say the personalized service.

On the other hand, the prevalence of photo capturing devices, together with the advent of media sharing services like Flickr [2] and Picasa [4], have led to the appearance of voluminous digital photos with text tags, timestamp and geographical references on the Internet. Different from other community-contributed multimedia data, these photos connect geography, time and visual information together and provide a unique data source to discover patterns and knowledge of our human society.

Sharing geo-tagged photos has been a common activity in social networks. Here have been many researchers contributed themselves to study the potential applications of these geo-tagged photos. However, the present raw photos cannot provide information as enough as complete GPS trajectories, the information such as the complete location information of travel route cannot be obtained due to the defects hidden in them. In this paper, we mainly aim at analyzing the large amounts of geo-photos by tracing people's trips and proposing a novel travel route restoring method to restore the raw geo-tagged photo route into complete GPS trajectory.

Our study is to work from the following steps: 1. Select the suitable geo-tagged photo trajectory by using mobility entropy; 2. Apply the density-based spatial clustering arithmetic to cluster all the photos and obtain the hot regions. Additionally, we rank the hot regions by computing the Interest Measure ratio; 3. Estimate the pause time on taking photos and redefine the photo trajectory; 4. Apply HSMM to transform the photo route into the significant point sequence; 5. Restore the significant point sequence and obtain the final complete GPS trajectory.

The structure of this paper is as follows: Section 2 reviews related works on the research of trajectories. Section 3 analyzes the geo-tagged photo trajectory and makes some important definitions. Section 4 applies the HSMM and Mean Value method to restore the photo trajectory and obtains the complete GPS trajectory. Section 5 proposes an evaluation method and presents the experiment results. Finally, we provide a conclusion and offer an outlook of future work in Section 6.

2. Related Work

During the recent years, motivated by the convenience of collecting trajectory data, a branch of research has been performed based on individual location history, these locations including GPS trajectories, geo-tagged photo routes and etc.

2.1 Trajectory Mining

Before the prevalence of GPS, most of trajectory mining is developed based on the text. In such days, it is inconvenient to record their trajectories with text, and a majority of trajectoriy records are travel blogs. [7] extracted typical visitor's travel routes by using a sequential pattern mining method. [8] mined location-representative tags from travelogues and then uses such tags to retrieve web images. [9] proposed a method to mine the association rules between locations, time periods, types of experiences out of blog entries. [10] established a model to identify the geographical locations of page contents to harvest city sight photos from Web blogs. However, the travel blog can only provide overall planning service but not the personalized route service, so it is still difficult to plan the travel with more detail.

Recent years have witnessed great prosperity in GPS trajectory service. Discovering, extracting, and summarizing knowledge from these data enable us to have more understandings about the world. More researchers begin to contribute themselves to study the GPS trajectories. [11] and [12] mined the interesting locations and classical travel sequences in a given geospatial region based on multiple users' GPS trajectories. [13] described a predestination method to predict where a driver is going as a trip progresses by using a history of a driver's destinations and the data about driving behaviors. [14] and [15] proposed an approach based on supervised learning to automatically infer transportation mode from raw GPS data. [16] studied a personalized friend and location recommender for the geographical information systems (GIS) on the Web. Obviously, the GPS trajectory can provide more accurate location information and travel service, but it is difficult to store and process the mass data. So it is necessary to find a method to resolve this problem.

2.2 Geo-tagged Photo Mining

As the rapid development of camera and mobile phone, sharing geo-tagged photos has been a common activity in social networks. Many researchers have contributed themselves to mine the Geo-tagged photo trajectories. [17], [18] and [20] proposed a method to detect people's frequent trip patterns, typical sequences of visited cities and etc. [19] employed the state-of-the-art object recognition technique to mine representative photos of the given concept for representative local regions from a large-scale unorganized collection of consumer-generated geo-tagged photos. [21] proposed a travel route recommendation method that makes use of the photographers' histories. [22] studied the leverage existing travel clues recovered from 20 million geo-tagged photos to suggest customized travel. [23] proposed a travel path search system based on geo-tagged photos to facilitate tourists' trip planning. [24] proposed to conduct personalized travel recommendation by further considering specific user profiles or attributes.

Obviously, the photos used in the researches are obtained from previous travelers` actual experience, and then mining photo trajectory patterns are beneficial for practical applications and services, such as travel advisory. Additionally, it is not necessary to store mass data with geographic information. However, it is difficult to obtain the information such as the significant regions travelers have gone, the completed location information and other information we can obtained from completed GPS trajectories but cannot obtained from Geo-tagged photo trajectory.

The completed GPS trajectory and Geo-tagged trajectory have their own advantages and disadvantages. Though there have been many researches on GPS trajectory mining Geo-tagged photo mining, no research consider the relationship between GPS trajectory and

Geo-tagged photo trajectory. In this paper, we intend to restore the Geo-tagged photo trajectory into completed GPS trajectory, then the disadvantages of both two kinds of trajectory can be complemented, and the advantages of them can be strengthened.

3. Data Preprocess

The data we used are collected from the web of Flickr. We choose the city of Beijing for collecting geo-tagged photo routes. **Fig1** shows the geo-tagged photos of Beijing in 2011.

3.1 Data Preprocessing

3.1.1 Definition

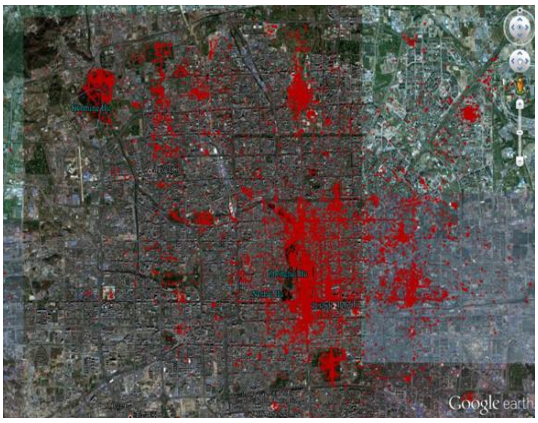


Fig. 1. Geo-tagged photo points in Beijing

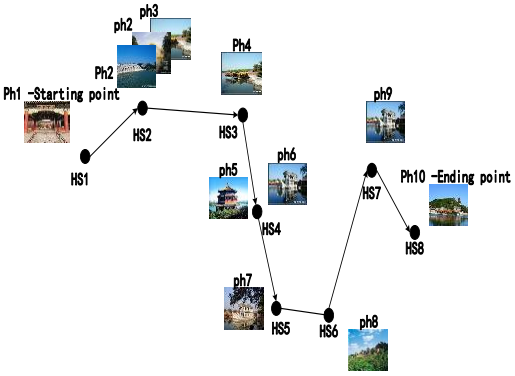


Fig. 2. An example of tourist photo rout

Given a set of GPS-tagged photos $PH = \{ph\}$ within a tour destination, we build the database of travel paths. Here, the tourist travel movement is modeled at a daily basis. According to photographer ID, we organize photos of each photographer in a day in a chronical sequence. **Fig. 2** shows a complete travel path, HS means "Hot Spot". Just as **Fig. 2**. illustrates, the definition of Geo-tagged Photo trajectory is defined as follows:

Definition 3.1 [Geo-tagged Photo trajectory]

The Geo-tagged Photo trajectory is defined as a finite chronological sequence of photo points, it is denoted as $Tra_{ph} = \{ph_1, ph_2, ..., ph_{np}\}$, where $ph_i = (l_i, t_i)$, $|t_{np} - t_1| < 24h$, t_i is the time stamp of photo i with $t_i < t_{i+1}$. np is the number of photos in the trajectory Tra_{ph}. □

Generally, the location l_i refers to the location where the *ith* photo was taken, but sometimes marks the location of the photographed object, l_i is represented as (lat_i, lon_i) which is the latitude and longitude of ph_i. The time t_i generally marks the time captured the object, but occasionally refers to the time the photo was uploaded to Flickr. Assuming that $|t_{np} - t_1| < 24h$, that means we define a complete travel trajectory path with travel time shorter than 24 hours (a day).

3.1.2 Select Tourist Trajectories

Generally, these spatial-temporal trajectories can be classified into tourist and non-tourist trails. The premise for classification is that the mobile nature of sightseeing renders the photos

of a true tourist to be spread over a large spatial extend within the tour, and there should be not so much repeated geo information.

From the probabilistic perspective, the mobility complexity leads to a geospatial distribution of photos with reasonably high entropy. Let $\Pr(l)$ be the geospatial density of photos with geospatial coordinates l pertaining to the user. The mobility entropy $H(Tra_{ph})$ of a tourist trajectory $Tra_{ph}=\{(l_1,t_1),...,(l_{np},t_{np})\}$ is computed as follows:

$$H(Tra_{ph}) = \sum_{i=1}^{np} \Pr(l_i) \log_2 \Pr(l_i) \tag{1}$$

where $\Pr_i(l_i)$ is the discrete geospatial distribution of photos in l_i, it is estimated by the counts of photos in l_i. To discriminate photo trails, we empirically set a mobility entropy threshold ε_{H}. If $H(Tra_{ph}) > \varepsilon_{\mathrm{H}}$, the photo trail PH can be seen as the tourist trajectory.

3.1.3 Pause Time

The time of taking photograph is important in the analysis of travel pattern and the establishment of HSMM model. The geo-tagged photos can be seen as the Interest of Points (POI) in the life or travel of users, and then we estimate the pause time of each POI by using the exponential smoothing method.

Denote $dis(x,y)$ as the earth distance function of points x and y, $T(x,y)$ as the travel time between x and y, then $v(ph_i, ph_{i+1}) = dis(ph_i, ph_{i+1}) / T(ph_i, ph_{i+1})$ is the travel speed of photographer/up loader. We save the notation of $v_i = v(ph_i, ph_{i+1})$, $dis_i = dis(ph_i, ph_{i+1})$, $T_i = T(ph_i, ph_{i+1})$. The pause time in the ith photo can be estimated as

$$pt_i = T_i - dis_i / v_i^* \tag{2}$$

where v_i^* is the average speed of the user, computed as exponentially weighted moving average:

$$v_i^* = \lambda v_i + (1-\lambda) v_{i-1}^*$$

The initial value of average speed can be set as v_1, that is $v_1^* = v_1$. The critical problem of exponential smoothing method is to determine a suitable λ. When $\lambda = 0$, $v_i^* = v_{i-1}^*$, when $\lambda = 1$, $v_i^* = v_i$. Generally, if the speed changes greatly, it will be better to choose a bigger λ, when there is a smooth speed sequence, a smaller λ is better. In the reality, users tend to walk with different speed. If he is interested with some places, he tends to walk with a slower speed, however, if he is on the road to the interesting spots, he may tend to walk with faster speed. we believe that most of travelers change the speed greatly. In this paper we set $\lambda = 0.75$.

3.2 Hot Spots

3.2.1 DBSCAN

In the travel, people prefer to describe the location with Landmark Buildings or spots. However, it is difficult to extract the spot information with simple single geo-information. In this paper, we use the clustering method to help users to recognize the hot spot (HS) and find that the algorithm of density-based spatial clustering of applications with noise (DBSCAN) performs the best among those clustering methods. A significant point sequence is mada up by HS points and geo-tagged photo points, just as **Fig. 2** shows. In the DBSCAN algorithm, the density of cluster (ρ) is determined by two parameters: ε (Eps) and the minimum number of points required to form a cluster (minPts). We present the clustering results in **Table 1**.

Table 1. The clustering information according to different density

Density of Cluster		Cluster		Spot				
ε (km)	MinPts	Class Number	NPR	Summer Palace	Tiananmen Square	Great Wall	Olympic Park	MIN
0.5	10	50	63.70%	2350	327	311	197	11
	20	21	71.88%	2069	300	124	158	22
	30	14	76.30%	1207	639	297	156	34
1	10	73	46.29%	2897	590	515	416	10
	20	22	58.60%	2446	535	406	314	18
	30	15	63.02%	2349	353	272	167	30
3	10	54	22.32%	6114	715	778	504	10
	20	30	30.94%	5075	689	750	572	19
	30	19	37.06%	4949	685	736	391	33

In **Table 1**, NPR is the noise point rate, MIN is the minimum number of photos in spots. From the table, we can have the following conclusion: on one hand, for the fixed MinPts, as ε increases, the class increases slightly, the NPR decreases rapidly, and the number of photos in each spot increases; on the other hand, for the fixed ε, as MinPts increases, the class number decreases, the NPR increases, and the photos in spots are less.

In this paper, let MinPts=20, $\varepsilon = 3km$, and the class number is 30. The parameters are chosen according to the following two principles:

i Firstly, We tend to choose the parameters which lead to smaller noise point rate, then $\varepsilon = 3km$ and MinPts=20 or MinPts=10.

ii Secondly, according to the reality of Beijing, we prefer to the parameters which lead to reasonable number of spots. Obviously, it is better to choose $\varepsilon = 3km$ and MinPts=20.

3.2.2 Interest Measure of Hot Regions

People take photos according to their subjective consciousness. Though there exists differences among their interests, they tend to take more photos in the beautiful spots and fewer photos in the less beautiful spots. We rank the spots according to the interest measure (IM), and IM can be computed by considering the number of photos and travelers in such spots.

$$\mathrm{IM}j = \left(\frac{N_j(ph)_{\rho 0}}{\sum_j N_j(ph)_{\rho 0}} + \frac{N_j(U)_{\rho 0}}{\sum_j N_j(U)_{\rho 0}} \right) \times \frac{1}{NPR_{\rho 0}} \tag{3}$$

Assuming that $\rho = \rho_0$, ρ_0 is the density parameter we choose. In (3), IM_j is the interest measure in the *jth* spot, $j = 1,2,\cdots n_s$, where n_s is the number of spots . $N_j(ph)_{\rho 0}$ is the number of photos in the *jth* spot, $N_j(U)_{\rho 0}$ is the number of users, $NPR_{\rho 0}$ is the corresponding NPR when $\rho = \rho_0$.

For a fixed ρ_0, $NPR_{\rho 0}$ cannot influent the relative value of interest measure of each spot. However, lower noise point rate may lead to more points in every spot, then there are more efficient points to compute the IM value. So by comparing with higher NPR value, lower NPR can lead to the more suitable IM value.

Considering the NPR and the computing complexity, we let $\varepsilon = 3km$ and MinPts=20 in the rest of paper, then $n_s = 30$.

4. Route Restoring

4.1 Model of HSMM

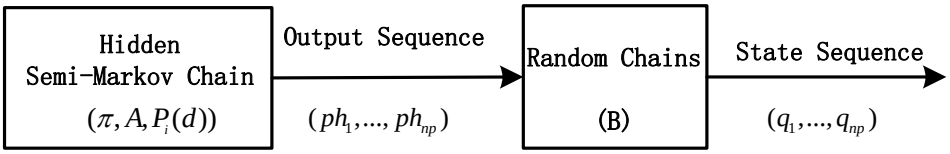


Fig. 3. The parameters in the HSMM.

The parameter set of HSMM is denoted as $\mu = (ns, no, \pi, A, B, P)$. **Fig. 3** illustrates the parameters of HSMM in detail. We aims at finding the most likely underlying state sequence $Q^* = \{q_1, q_2, ..., q_{nq}\}$ of geo-tagged photo trajectory.

(1) *ns* is the number of states in the state set S, $S = \{S_1, S_2, ..., S_{ns}\}$. We have set $ns = 30$ in this paper.

(2) *no* is the number of observation sequence (output sequence), $\mathrm{O} = \{O_1, O_2, \cdots, O_{no}\}$. We see the geo-tagged photo as output sequence, then $\mathrm{O} = \mathrm{Ph} = \{ph_1, ph_2, ..., ph_{np}\}$, and $no = np$.

(3) π is the initial state probability. $\pi = \{\pi_1, \pi_2, \cdots, \pi_{\mathrm{ns}}\}$, where π_i is the probability when the initial state is i, $\pi_i := P(q_1 = i)$ with $\sum \pi_i = 1$.

(4) $A = [a_{ij}]_{ns \times ns}$ is the state transition probability matrix (STPM). Considering the first order Markov Chain, the current state q_i only depends on the previous state q_{i-1}, that is $a_{ij} = P(q_i = S_j \mid q_{i-1} = S_i)$. STPM can be obtained by counting transitions.

(5) $B=[b_j(k)]_{np\times nq}$ is the state output probability distribution.

(6) $P=\{P_1(d),P(d)_2,...,P_{ns}(d)\}$ is the set of state duration probability. $P_i(d)$ is the probability of duration time the user stayed in state i, $P_i(d):=P(q_{t+d+1}\neq j,q_{t+d}=j,\ldots,q_{t+2}=j\,|\,q_{t+1}=j,q_t=i)$, where $1\le d\le D$, D is the longest duration time of users, we assume that $D<24h$.

Assuming that the *ith* state output and duration distribution are given by Gaussian density function with mean μ_1 and μ_2 and variance σ_1 and σ_2 respectively.

$$b_j(k)=N(\mathrm{O}\,|\,\mu_1,\sigma_1)$$
$$P_i(d)=N(d\,|\,\mu_2,\sigma_2)$$

As a result, the parameter set of HSMM λ can be simplified as $\lambda=(\pi,A,\mu_1,\sigma_1,\mu_2,\sigma_2)$.

Travelers would like to prefer the spot with higher interest measure and the short distance between current locations to that spot. Then the initial state probability is computed as

$$\pi_i=\frac{\mathrm{IM}_i}{d_i\sum_i\frac{\mathrm{IM}_i}{d_i}}$$

IM_i is the interest measure of spot i, d_i is the distance from the first geo-tagged photo to the *ith* spot.

We have known that $a_{ij}:=P(q_{t+1}=j\,|\,q_{t+1}\neq i,q_t=i)$, then the transition probability is computed by counting the transition frequency, $a_{ij}=N_{ij}/\sum_k N_{kj}$, where $N_{ij}=2^{n_{ij}}$, n_{ij} is the transition frequency from state i to j.

The duration time is just the pause time estimated in the section 3.1.3. Additionally, μ_2 and σ_2 are estimated with mean value $\overline{pt}$ and $\sum_{i=1}^{ni}(pt_i-\overline{pt})^2/(n_i-1)$, *ni* is the number of users who have taken photos in *ith* spot.

Assuming that the state output probability also depends on IM and the distance between current locations to the spot. Then $b_j(k)=N(\mathrm{IM}_j/d_{jk}\,|\,\mu_1,\sigma_1)$. μ_1 and σ_1 can be estimated with the same method of duration time distribution.

Viterbi algorithm is used to estimate the most likely underlying state sequence $Q^*=\{q_1,q_2,...,q_{nq}\}$.

4.2 Discovering SP Sequence

We see all the geo-tagged photos and the estimated state sequence are significant points (SPs) for a traveler. **Fig. 4** shows a SP sequence. However, there is no time stamp of $Q^* = \{q_1, q_2, ..., q_{nq}\}$. In this section, we intend to rearrange these points with time sequence.

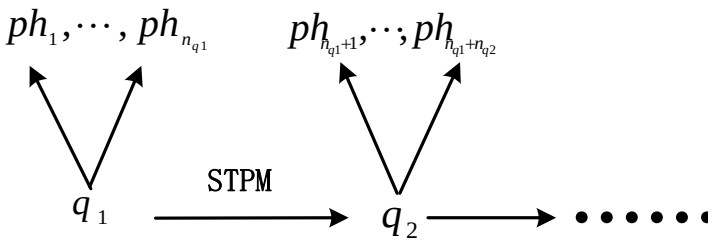


Fig. 4. The basic SR sequence for a user.

Take the geo-tagged photo trajectory $\mathrm{Ph} = \{ph_1, ph_2, ..., ph_{np}\}$ and the estimated state sequence $Q^* = \{q_1, q_2, ..., q_{nq}\}$ as example. Obviously, there exists two possibilities of SP sequence in the first state, with the same theory, there are 2^{nq} possibilities of SR sequences in all the states. We use the following two principles to discover the most likely SR sequence for the traveler.

Shortest distance principle:

We choose the sequence with the shortest travel length in the state q_i.

Main direction principle:

Denote that $\vec{v}_i = (X_i, Y_i) = (lat_{i+1} - lat_i, lon_{i+1} - lon_i)$. The sigh function of $\vec{v}_i$ can be defined as $\mathrm{sgn}(\vec{v}_i) = (\mathrm{sgn}(X_i), \mathrm{sgn}(Y_i)) = (I_{lat}, I_{lon})$. To account for the main direction of geo-tagged photo trajectory, we summarize $\mathrm{sgn}(\vec{v}_i)$, and consider the summarization $\sum_i \mathrm{sgn}(\vec{v}_i) = (\sum_i I_{lati}, \sum_i I_{loni}) = (I_{lat}, I_{lon})$ as the main direction. If $|I_{lat}| > |I_{lon}|$, we choose the sequence with the same sign of latitude. If $|I_{lat}| = |I_{lon}|$, Repeat this process with the first half part of the geo-tagged photos. The shortest distance principle is the first choice when there is a contradiction.

We take **Fig. 5** as example to illustrate the shortest distance principle and main direction principle. In the first state q_1, there are three photos points and two kinds of SP sequences (the blue one and the green one). It is easy to see that, $dis(ph_1, ph_2) + dis(ph_2, q_1) + dis(q_1, ph_3) < dis(ph_1, q_1) + dis(q_1, ph_2) + dis(ph_2, ph_3)$, then we choose the blue sequence. The traveler may first went to q_1, and then take the photo ph_2. Additionally, $\sum \mathrm{sgn}(\vec{v}_i) = (4,2)$, 4>2, $\vec{v}_i(ph_2, q_1) = (-1,1)$, $\vec{v}_i(q_1, ph_2) = (1,-1)$, the sign of 1 in $\vec{v}_i(q_1, ph_2)$ is same with 4, then we choose the blue sequence. The final sequence of SP is $ph_1 \to q_1 \to ph_2 \to ph_3 \to ph_4 \to q_2 \to q_3 \to ph_5$.

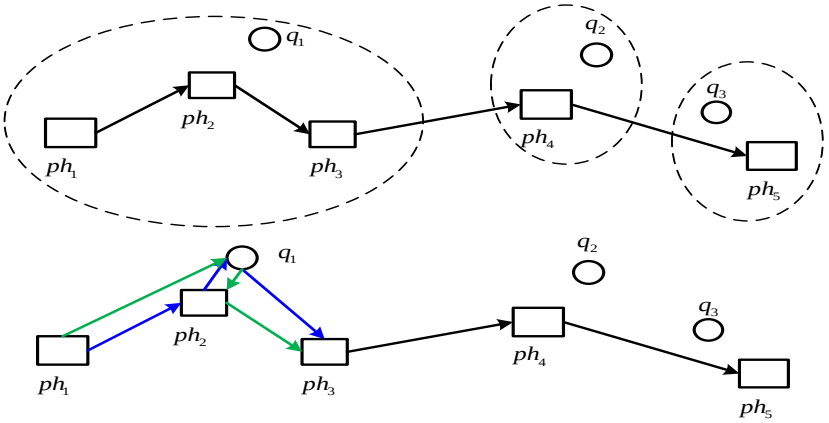


Fig. 5. Output state sequence and geo-tagged photo sequence.

4.3 Discovering GPS Trajectory

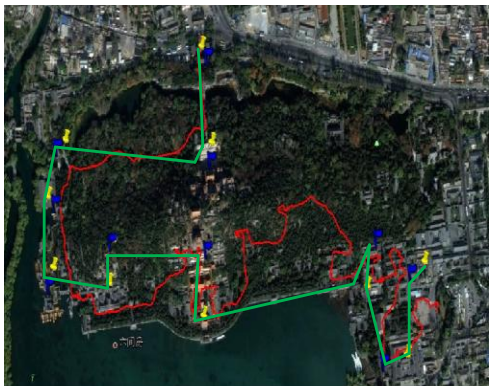


Fig. 6. Raw GPS trajectory and the SP sequence obtained from HSMM.

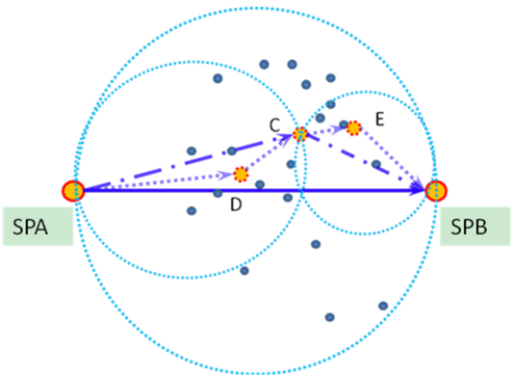


Fig. 7. Discovering GPS route between SPA and SPB.

In **Fig. 6**, the blue flags are output state obtained from HSMM, the yellow ones are geo-tagged photo uploaded by a user, the green line is the SP route, and the red line is a raw GPS trajectory covers all the output states. From the figure, we can see that it is convenient to obtain the spot information, but people cannot understand the trajectory clearly and further make perfect travel plan in advance, there is still big gap between SP sequence and the complete GPS trajectory. We propose a Mean Value (MV) method to obtain the complete GPS trajectory from SP sequence.

The Mean Value (MV) method is established based on the large amounts of history photo data uploaded by various travelers, because the travelers would like to select the hot trajectory and hot spot. Then we can get a traveler's GPS trajectory by considering other travelers' mobility rules.

We take two significant points as example to illustrate the mean value method. Just as **Fig. 7** shows, SPA and SPB are two significant points in the SP sequence. In our approach we first connect SPA and SPB, then there is a circle with radius of $1/2 \cdot dis(\text{SPA,SPB})$. We then compute the mean value of all geo-tagged photos in this circle, if $1/2 \cdot dis(\text{SPA}, \text{SPB}) < \varepsilon$ and the number of photos in the circle less than MinPts ($n_{AB} < MinPts$), then stop. Otherwise,

denote the mean value as a GPS point C in the trajectory, and the route (SPA, SPB) becomes (SPA, C, SPB). Repeat this process in route (SPA, C) and (C, SPB) until the radius of circle less than ε and the number of photos in the circle less than MinPts. With the same process of (SPA, SPB), the routes between other significant points can be restored. Finally, we obtain the complete GPS trajectory.

It has been proved that the complexity of Viterbi algorithm used in HSMM is $O(N^2T)$, N is the number of spots we found in a city, T is the number of Geo-tagged photos. The complexity of algorithm we used in 4.2 is 2^N. The complexity of Mean Value algorithm we used in 4.3 is N. Then the complexity of our trajectory restoring algorithm is $O(N^2T)$.

5. Experiment

We intend to test the restoring method from two parts. On one hand, we test the reality of restored GPS trajectory, to show whether it is agree with the travelers' mobility regularity. On the other hand, we test its accuracy, to show whether the restored GPS trajectories are conform to travelers' usual GPS trajectories.

5.1 Power Law

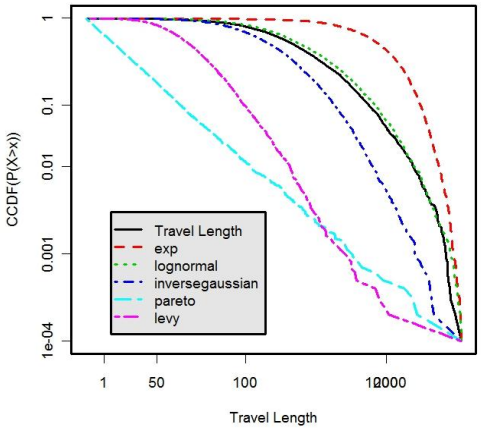


Fig. 8. CCDF plot of travel length.

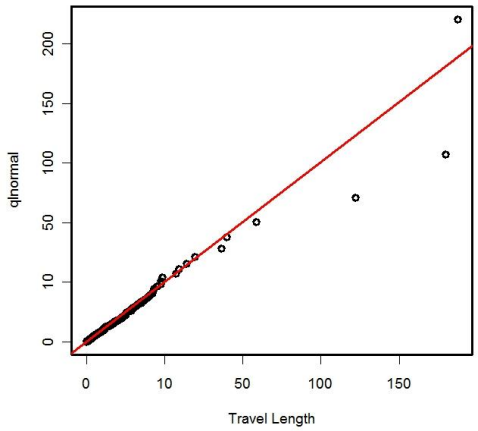


Fig. 9. QQ-plot of travel length and log-normal distribution.

[25] and [26] have reported that the mobility patterns of wild capuchin monkeys are not random walks, and people also show the similar recurrence properties. **Fig 8** shows the CCDF (complementary cumulative density function) of travel length from all the complete GPS trajectories obtained from HSMM and MV method. Travel length is the distance between two adjacent GPS points. CCDF is known to show the tail patterns of a distribution better than log-log binned PDF plots. We apply Maximum Likelihood Estimation (MLE) to fit five known distributions, exponential, log-normal, inverse Gaussian, truncated Pareto and Levy distribution of the CCDF. We observe that log-normal distribution performed over than other distributions in all cases, which is a rule of thumb for power-law distribution.

Fig. 9 shows the QQ-plot of travel length and log-normal distribution. QQ-plot is a probability plot, it is a graphic method for comparing two distributions by plotting their quantiles against each other. If the two distributions being compared are similar, the points in the QQ-plot will approximately lie on a line. Obviously, **Fig. 9** shows that the travel length fitted log-normal distribution. The results in two figures show that the complete GPS trajectories obtained from HSMM and MV method are in accordance with travelers' mobility regularity.

5.2 Evaluation Method

Let Tra_{GPS} and Tra_{SP} be the raw GPS trajectory and trajectory obtained from HSMM and MV method respectively, $Tra_{GPS} = (p_1, p_2, ..., p_{n_G})$, $Tra_{SP} = (p'_1, p'_2, ..., p'_{n_{SP}})$. Both trajectories cover the same spots. $dif(\cdot)$ is the difference between two trajectories. We use the Leave One Out Cross Validation Ratio (LOOCV Ratio) to test the performance of our trajectory restoring method,

$$\text{LOOCV}(Tra_{GPS}, Tra_{SP}) = \frac{PRSD(Tra_{GPS}, Tra_{SP})}{SoD(Tra_{GPS}, Tra_{SP})} = \frac{\sum_{i=1}^{np} dif(Tra_{GPS}, Tra_{SP(i)})}{np \times dif(Tra_{GPS}, Tra_{SP})} \tag{4}$$

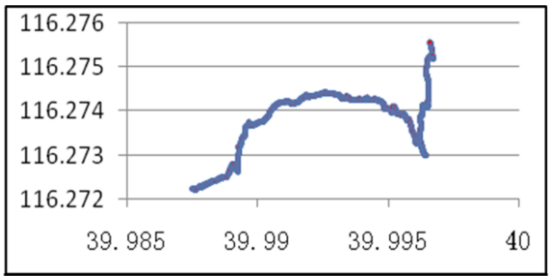
PRSD is the predicted sum of differences of two trajectories, SoD is the sum of Differences. $Tra_{SP(i)}$ is the GPS trajectory obtained from the restoring method when the *ith* geo-tagged photo is removed from the training data. $dif(Tra_{GPS}, Tra_{SR})$ is a function to illustrate the difference between two trajectories. Naturally, the points in Tra_{SP} are less than Tra_{GPS}, then $dif(Tra_{GPS}, Tra_{SP}) = \sum_{i=1}^{n_{SP}} d(p'_i, Tra_{GPS})$, where $d(p'_i, Tra_{GPS}) = dis(p'_i, p_k)$ is the function to compute the distance between point p'_i and trajectory Tra_{GPS}, $k = \arg\min_{0<j<n_{SP+1}} dis(p'_i, p_j)$, $j = 1, 2, ..., n_{SP}$.

Generally, LOOCV may lager than 1, the ratio can be used to estimate how accurately the GPS trajectory restoring model will perform in practice.

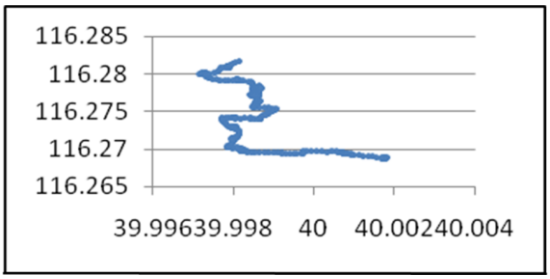
5.3 Main Results Obtained From Restoring Method

Fig. 10 shows the results of two completed trajectories of tra_{ph1} and tra_{ph2} by applying HSMM and MV method. The pictures in the first row ((a1) and (a2)) are two GPS trajectories tra_1 and tra_2, these two trajectories are collected by the volunteers. In the tra_1, the GPS point number is 1188s, and the time span is one hour and 31 minutes. In the tra_2, the GPS point number is 457, and the time span is 46 minutes. The pictures in the second row ((b1) and (b2)) are two photo trajectories we selected in the trajectory set, additionally, these two trajectories covers the same spots of tra_1 and tra_2, in another words, the spots in tra_{ph1} and tra_1 are the same, tra_{ph2} and tra_2 have the same spots. The pictures in the third row ((c1) and (c2)) are two spot sequences obtained from HSMM algorithm. The pictures in the fourth row ((d1) and (d2)) are two completed trajectories obtained from the restoring method.

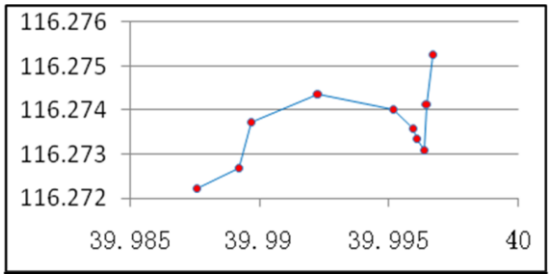
From **Fig. 10** we can have the following conclusion: Even there are only geo-tagged photos, we can still obtain the complete GPS trajectory of travelers, and furthermore, both the raw trajectory and restored trajectory are so similar. Then it is more convenient to make travel plans in advance.



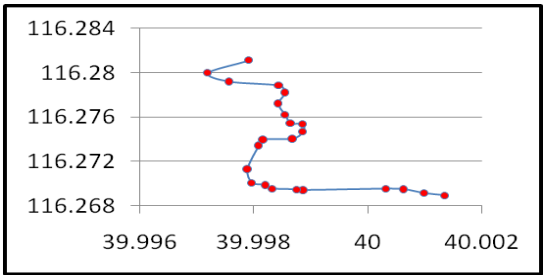
(a1) Raw GPS trajectory tra_1



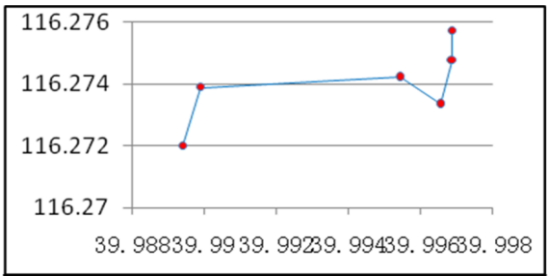
(a2) Raw GPS trajectory tra_2



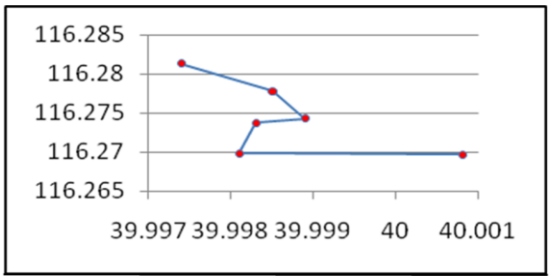
(b1) Phtoto trajectory tra_{ph1} covers the same spots in tra_1



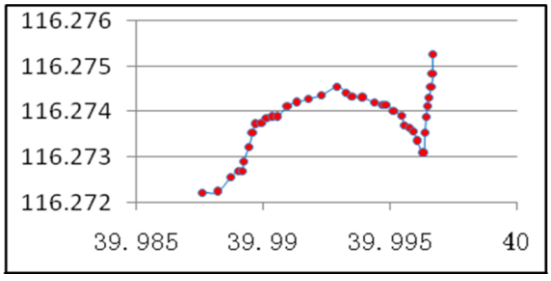
(b2) Phtoto trajectory tra_{ph2} covers thesame spots in tra_2



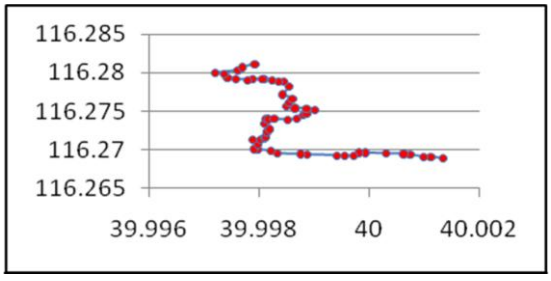
(c1) The spots sequency of tra_{ph1} restored from HSMM



(c2) The spots sequency of tra_{ph2} restored from HSMM



(d1) The completed trajectory of tra_{ph1} restored from HSMM and MV method



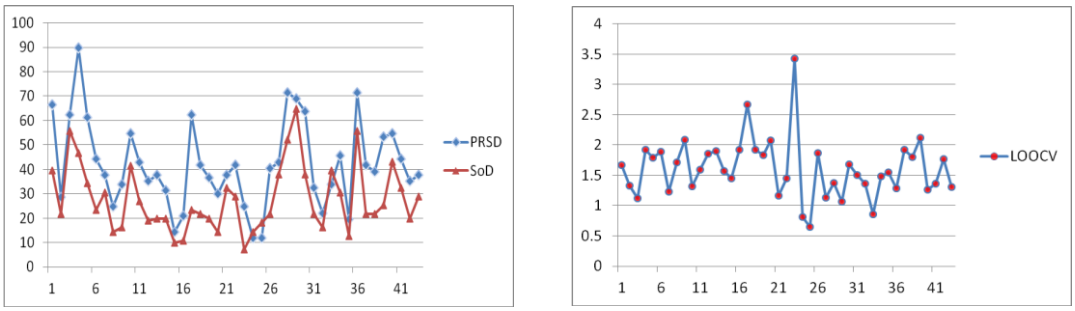
(d2) The completed trajectory of tra_{ph2} restored from HSMM and MV method

Fig. 10. The results of completed trajectory obtained from HSMM and MV method.

In the experiment, we chose 43 pairs of trajectories to test the performance of restoring method. In each pair, there is a photo trajectory and a raw GPS trajectory, additionally, they

have the same spots. In another words, the person generated the GPS trajectory and the photographer generated the photo trajectory have traveled in the same spots. We first restored the photo trajectories into completed GPS trajectories, and then computed the differences between the completed GPS trajectories and its corresponding raw GPS trajectories with the method mentioned in Section 5.1.

Two figures in **Fig. 11** show the results of PRSD, SoD, and LOOCV ratio from 43 pairs of trajectories when $\lambda = 0.75$. Both trajectories in each pair cover the same spots. The average value of LOOCV Ratio is 1.61. From the figure we can see that the HSMM and MV method performs well in restoring geo-tagged photo trajectory.



(a). The results of PRSD and SoD (b). The result of LOOCV

Fig. 11. The result of PRSD, SoD, and LOOCV ratio from 43 pairs of trajectories.

6. Conclusion

The GPS-tagged photos available on the Internet implicitly provide spatial-temporal movement trajectories of their photographers, additionally, they are obtained from previous travelers` actual experiences and the cost of collecting detailed travel data is formidable. So the analysis on tourist mobility is important to tourism bureaucracy and industries. We first built a statistically reliable tourist movement geo-tagged photo route database from GPS-tagged photos, by utilizing an entropy-based mobility measure and estimating the pause time with exponential smoothing method. We then obtain the hot spots by using DBSCAN algorithm, and compute the interest measure of each spot by considering the number of photos and travelers in it. Next, we establish the HSMM to obtain the output state sequence, combined with the geo-tagged photo sequence, these points compose a significant point sequence. Finally, the complete GPS trajectory is obtained by applying the Mean Value method. In the experiment, we test the restoring method from two parts: the reality of restored GPS trajectory and its accuracy. The results show that the proposed approach can deliver promising results. One of our future work is to leverage the GPS-tagged photos to analyze the tourist travel behavior.

References

[1] Yahoo travel. http://travel.yahoo.com/
[2] Gps Track log route exchange Forum. http://www.gpsxchange.com/
[3] Flickr. http://www.flickr.com/
[4] Picasa. http://picasa.google.com/
[5] C. H. Lo, W. C. Peng, C. W. Chen, T. Y. Lin and C. S. Lin, “CarWeb: A Traffic Data Collection Platform,” in *Proc. of 9th Int. Conf. on Mobile Data Management*, pp. 221-222, April 27-30, 2008. Article (CrossRef Link)

[6] Y. M. Chang, L. Y. Wei, C. S. Lin, C. H. Jung, W. C.Peng and I. H. Chen, "Exploring GPS Data for Traffic Status Estimation," in *Proc. Of 10th Int. Conf. on Mobile Data Management*, pp. 369-370, May 18-20, 2009. Article (CrossRef Link)

[7] H. Kori, S. Hori, T. Tezuka and K. Tanaka, "Automatic Generation of Multimedia Tour Guide from Local Blogs," in *Proc. of 13th Int.Multimedia Modeling Conference*, pp. 690-699, January 9-12, 2007. Article (CrossRef Link)

[8] Q. Hao, R. Cai, X.J. Wang, J.M. Yang, Y. Pang, L. Zhang, "Generating Location Overviews with Images and Tags by Mining User-Generated Travelogues," in *Proc. of 17th ACM international conference on Multimedia*, pp. 801-804, October 19-24, 2009. Article (CrossRef Link)

[9] T. Kurashima, T. Tezuka and K. Tanaka, "Mining and Visualizing Local Experiences from Blog Entries," in *Proc. of 17th International Conference on Database and Expert Systems Applications*, pp. 213-222, September 4 – 8, 2006. Article (CrossRef Link)

[10] R. Ji,, X. Xie, H. Yao, and W.Y. Ma, "Mining City Landmarks from Blogs by Graph Modeling," in *Proc. of 17th ACM Multimedia*, pp. 105-114, October 19-23, 2009. Article (CrossRef Link)

[11] Y. Zheng, L. Zhang, X. Xie and W.Y. Ma, "Mining interesting locations and travel sequences from GPS trajectories," in *Proc of Int. conf. on 18th World Wild Web*, pp. 791-800, April 20-24, 2009. Article (CrossRef Link)

[12] D. Ashbrook and T. Starner, "Using GPS to Learn Significant Locations and Predict Movement across Multiple Users," *Personal and Ubiquitous Computing*, vol.7, no.5, pp. 275–286, 2003. Article (CrossRef Link)

[13] J. Krumm and E. Horvitz, "Predestination: Inferring Destinations from Partial Trajectories," in *Proc. of Int. Conf. on 8th Ubiquitous Computing*, pp. 243–260, May 7-10, 2006. Article (CrossRef Link)

[14] Y. Zheng, L. Liu, L. Wang, X. Xie, "Learning Transportation Modes from Raw GPS Data for Geographic Application on the Web," in *Proc. of Int. conf. on 17th World Wild Web*, pp. 247-256, April, 21-25, 2008. Article (CrossRef Link)

[15] Y. Zheng, Q. N. Li, Y. Chen and X. Xie, "Understanding Mobility Based on GPS Data," in *Proc. of 10th UbiComp*, pp. 312-321, September 21 - 24, 2008. Article (CrossRef Link)

[16] Y. Zheng, L. Zhang and X. Xie, "Recommending friends and locations based on individual location history," To appear in *ACM Transaction on the Web*, vol.5, issue.1, article no.5, 2011. Article (CrossRef Link)

[17] Y. Arase, X. Xie, T. Hara and S.Nishio, "Mining people's trips from large scale geo-tagged photos," in *Proc. of 18th int. conf. on Multimedia* , pp. 133-142, October 25-29, 2010. Article (CrossRef Link)

[18] Y.Tao. Zheng, Y. Li, Z.Jun Zha and T.S. Chua, "Mining Travel Patterns from GPS-Tagged Photos," in *Proc. of 17th Int.Multimedia Modeling Conference*, pp. 262-272, January 5-7, 2011 Article (CrossRef Link)

[19] K. Yanai, B. Qiu, "Mining Cultural Differences from a Large Number of Geotagged Photos," in *Proc. of Int. conf. on 18th World Wild Web*. pp. 1173-1174, April 20-24, 2009. Article (CrossRef Link)

[20] S. Kisilevich, D. Keim, L. Rokach, "A novel approach to mining travel sequences using collections of geotagged photos," in *Proc. of 13th AGILE Int. Conf. on Geographic Information Science*, pp. 163-182, May 10-14, 2010. Article (CrossRef Link)

[21] T. Kurashima，T. Iwata，G. Irie and K. Fujimura, "Travel route recommendation using geotags in photo sharing sites," in *Proc. of the 19th ACM int. conf. on Information and knowledge management*, pp. 579-588, October 26 – 30, 2010. Article (CrossRef Link)

[22] X. Lu, C. Wang, J. M. Yang, Y. Pang, and L. Zhang, "Photo2Trip: Generating Travel Routes from Geo-Tagged Photos for Trip Planning," in *Proc. of 18th ACM Multimedia Int. Conf.*, pp. 143-152, October 25-29, 2010. Article (CrossRef Link)

[23] C. H. Wang, N. H. Yu, L. Zhang, "Trip Mining and Recommendation from Geo-tagged Photos," in *Proc. of the IEEE int. conf. on Multimedia and Expo Workshops*, pp. 540-545, July 9-13, 2012. Article (CrossRef Link)

[24] A. j. Cheng, Y. Y. Chen, Y. T. Huang, W. H. Hsu, H. Y. Mark Liao, "Personalized travel

recommendation by mining people attributes from community-contributed photos," in *Proc. of the 19th ACM international conference on Multimedia*, pp. 83-92, November 28 - December 01, 2011. Article (CrossRef Link)

[25] D. Boyer, M. C. Crofoot and P. D. Walsh, "Non-random walks in monkeys and humans," *Joural of the Royal Society Interface*, vol. 9 no. 70, pp 842-847, 2012. Article (CrossRef Link).

[26] I. Rhee, M. Shin, S. Hong, K. Lee, and S. Chong, "On the levy walk nature of human mobility," *IEEE/ACM Transactions on* networking, vol.19, no.3, pp. 630-643, June 2011. Article (CrossRef Link)

Guannan Wang is a PhD candidate statistics at Central South University, China. Her reaearch interests include context-aware pervasive computing, human mobility prediction, and trajectory data mining. Contact her at wgn1103@gmail.com.

Zhizhong Wang is a professor in statistics at Central South University, China. His research interest include empirical likelihood method, semi-parametric models, data mining, and applied statistics. Contact him at wzz8713761@163.com

Zhenmin Zhu is a professor in computer science at Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His research interest include context-aware pervasive computing. Contact him at zmzhu@ict.ac.cn.

Saiping Wen is a postgraduated studebt in computer science at Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His research interest include mechine learning and personalized recommendation. Contact him at wensaiping@ict.ac.cn