Route Restoring Based on Hidden Semi-Markov Model

Guannan Wang[2,3], Jian Ye[1,2], Zhenmin Zhu[1,2]

[1] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2] Beijing Key Laboratory of Mobile Computing and Pervasive Device, Beijing, China
[3] School of Mathematics Science and Computing Technology, Central South University, HuNan, China,
wgn1103@gmail.com, jye, zmzhu@ict.ac.cn

Abstract - 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 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.

Index Terms - Geo-tagged Photo, Hot spot, Travel route, Route restoring, Trajectory mining

1. Introduction

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 can not 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.

A. Trajectory mining

Recent years have witnessed great prosperity in GPS trajectory service. [1] and [2] mined the interesting locations and classical travel sequences in a given geospatial region based on multiple users' GPS trajectories. [3] 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. 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.

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. [4] and [5] mined the interesting locations and classical travel sequences in a given geospatial region based on multiple users' GPS trajectories. [6] 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.

B. 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. [7], [8] and [9] proposed a method to detect people's frequent trip patterns, typical sequences of visited cities and etc. [10] 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. Obviously, these 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.

3. Route Restoring

A. Model of HSMM

The parameter set of HSMM is denoted as $\mu = (ns, no, \pi, A, B, P)$. Fig1 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.
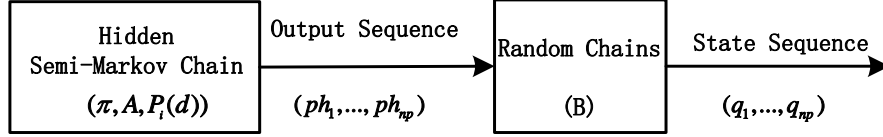
Fig.1 The parameters in the HSMM

(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), $O = \{O_1, O_2, L, O_{no}\}$. We see the geo-tagged photo as output sequence, then $O = Rh = \{ph_1, ph_2, ..., ph_{np}\}$, and $no = np$.

(3) π is the initial state probability. $\pi = \{\pi_1, \pi_2, \cdots, \pi_{ns}\}$, 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 $a_{ij} = P(q_i = S_j | q_{i-1} = S_i)$. 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, $P_i(d) := P(q_{t+d+1} \neq j, q_{t+d} = j, ..., q_{t+2} = j | q_{t+1} = j, q_t = i)$, where $1 \# d \quad D$, D is the longest duration time of users, we assume that $D < 24h$.

B. Discovering SP Sequence

We see all the geo-tagged photos and the estimated state sequence are significant points (SPs) for a traveler. 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.

Take the geo-tagged photo trajectory $Rh = \{ph_1, ph_2, ..., ph_{np}\}$ 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 $\vec{v}_i = (X_i, Y_i) = (lat_{i+1} - lat_i, lon_{i+1} - lon_i) \cdot$ The sigh function of $\vec{v}_i$ can be defined as

$sgn(\vec{v}_i) = (sgn(X_i), sgn(Y_i)) = (I_{lat}, I_{lon})$. To account for the main direction of geo-tagged photo trajectory, we summarize $\mathbf{sgn(\vec{v}_i)}$, and consider the summarization $\sum_i sgn(\vec{v}_i) = (\sum_i I_{lati}, \sum_i I_{loni}) = (I_{lat}, I_{lon})$ as the main direction. If $|I_{lat}| \triangleright |I_{lon}|$, 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. 2 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, $\sum sgn(\vec{v}_i) = (4,2)$, $4 > 2$, $\vec{v}_i(ph_2, q_1) = (-1,1)$, $\vec{v}_i(q_1, ph_2) = (1,-1)$, the sign of 1 in $\vec{v}_i(q_1, ph_2)$ is same with 4, then we choose the blue sequence. The finnal sequence of SP is $ph_1 \rightarrow q_1 \rightarrow ph_2 \rightarrow ph_3 \rightarrow ph_4 \rightarrow q_2 \rightarrow q_3 \rightarrow ph_5 \cdot$
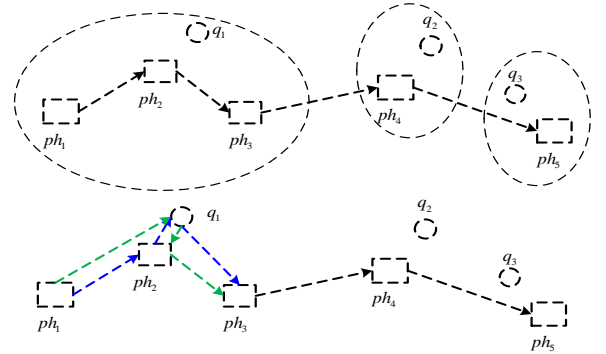

Fig. 2 Output state sequence and geo-tagged photo sequence

B. Discovering GPS trajectory.

In Fig 3, 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

121

between SP sequence and the complete GPS trajectory. We propose a Mean Value (MV) method to obtain the complete GPS trajectory from SP sequence.
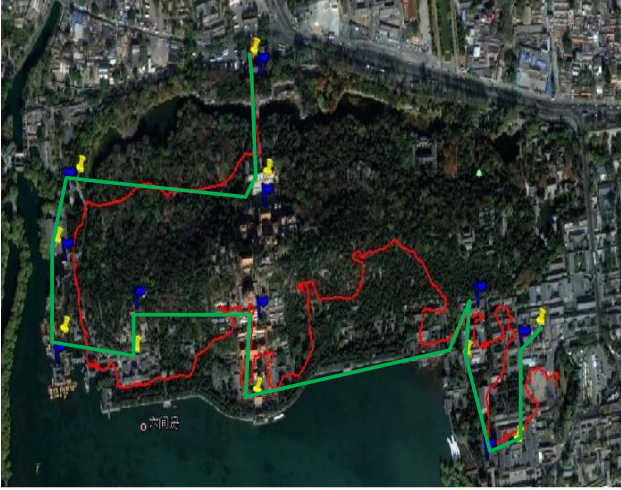

Fig 3**.** Raw GPS trajectory and the SP sequence obtained from HSMM.

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.

4. Experiment

We intend to test the reality of restored GPS trajectory, to show whether it is agree with the travelers' mobility regularity.

A. Power Law

[11] and [12] have reported that the mobility patterns of wild capuchin monkeys are not random walks, and people also show the similar recurrence properties. Fig. 4 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. 5 shows the QQ-plot of travel length and log-normal distribution, and there is a good performance. The results in two figures show that the complete GPS trajectories obtained from HSMM and MV method are in accordance with travelers' mobility regularity.
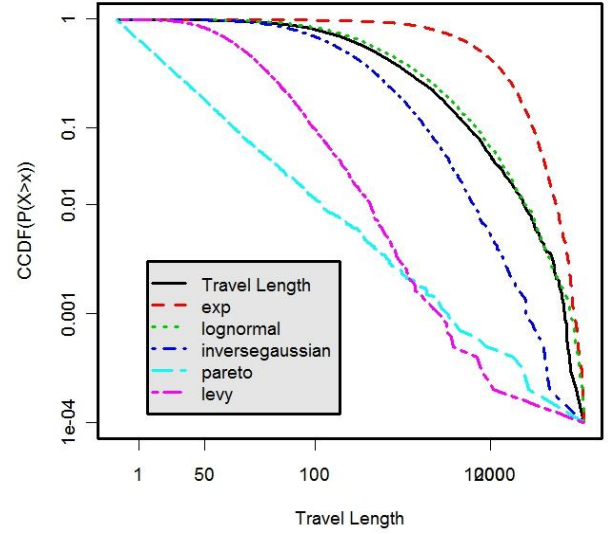

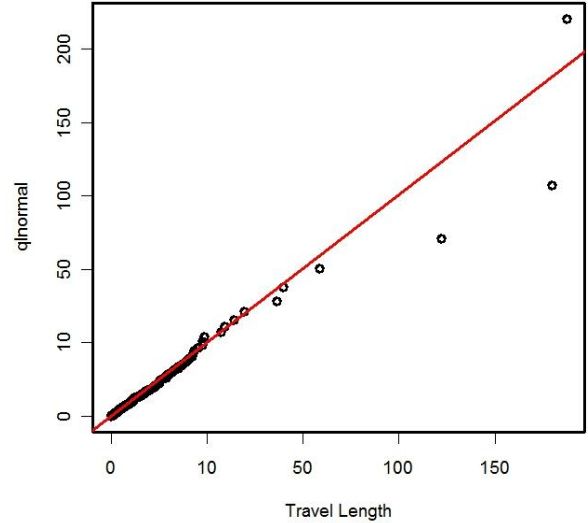Fig. 4 CCDF plot of travel length.

Fig. 5 QQ-plot of travel length and log-normal distribution.

B. Main Result

The second picture in Fig10 shows the GPS trajectory (Tra_{SP}) restored by HSMM and MV method. We choose a raw GPS trajectory (Tra_{GPS}) also covers the same spots with Tra_{SP}, and present the result in the first picture of Fig.6. From Fig. 6 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 two trajectories are so similar. Then it is more convenient to make travel plans in advance.
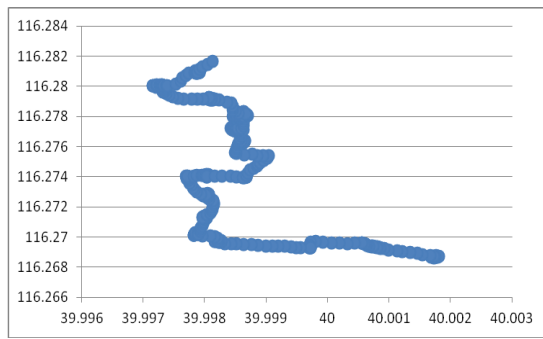
(a)   Raw GPS trajectory
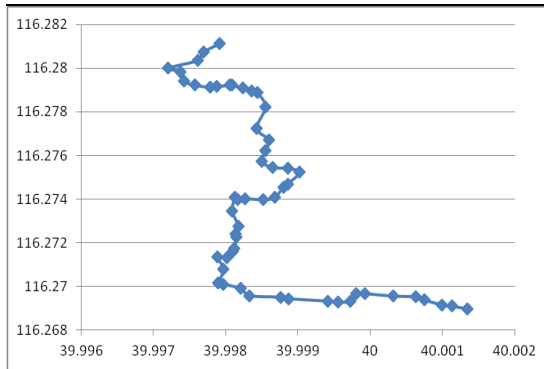
(b)   Complete GPS trajectory obtained from restoring method

Fig. 6.The raw GPS trajectory (Tra_{GPS}) and the GPS trajectory (Tra_{SP}) obtained from HSMM and Mean method.

5. 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 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] Y. Zheng, L. Zhang, X. Xie and W.Y. Ma, "Mining interesting locations and travel sequences from GPS trajectories", in *Pro. of Int. conf. on 18th World Wild Web*, pp 791-800, April 20-24, 2009.
[2] 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.
[3] J. Krumm and E. Horvitz, "Predestination: Inferring Destinations from Partial Trajectories", in *Proc. Int. Conf. on 8th Ubiquitous Computing*, pp.243–260, May 7-10, 2006.
[4] Y. Zheng, L. Zhang, X. Xie and W.Y. Ma, "Mining interesting locations and travel sequences from GPS trajectories", in *Pro. of Int. conf. on 18th World Wild Web*, pp 791-800, April 20-24, 2009.
[5] 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.
[6] J. Krumm and E. Horvitz, "Predestination: Inferring Destinations from Partial Trajectories", in *Proc. Int. Conf. on 8th Ubiquitous Computing*, pp.243–260, May 7-10, 2006.
[7] Y. Arase, X. Xie, T. Hara and S.Nishio, "Mining people's trips from large scale geo-tagged photos", in *Proc. of the 18th int. conf. on Multimedia* , pp. 133-142, October 25-29, 2010.
[8] 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
[9] 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.
[10] K. Yanai, B. Qiu, "Mining Cultural Differences from a Large Number of Geotagged Photos", in *Proc. of Int. conf. on 18th World Wild Web* WWW. pp. 1173-1174, April 20-24, 2009.
[11] Boyer, M. C. Crofoot and P. D. Walsh, "Non-random walks in monkeys and humans", J. R. Soc. Interface, vol.10, pp, 9: 842-847, 2011.
[12] Rhee, M. Shin, S. Hong, K. Lee, and S. Chong, "On the levy walk nature of human mobility". in Proc. of the27th Conference on Computer Communications, pp. 924-932, April 13-18, 2008.