An Uncertain Trajectory Modelling Method Based on Kernel Density Estimation

Yuan Cheng[1], RonghuaChi[2] and Yahong Wang[3]
{changuang7@sina.com[1], chironghua@126.com[2], wyh811619@126.com[3]}

College of Computer Science and Technology, Harbin University of Science and Technology[1],
College of Computer and Information Engineering, Heilongjiang University of Science and Technology[2]
School of Applied Science and Civil Engineering, Beijing Institue of Technology, Zhuhai[3]

Abstract. The accurate analysis of trajectories is of great significance for route selection, traffic status analysis, and urban traffic planning and so on. Existing researches lack effective methods for dealing with possible uncertainties in trajectories caused by objective enviroment and subjective intention etc. This work studies the method of constructing an uncertain model for the trajectories with the same starting point and end point based on kernel density estimation, to discover the distribution characteristics of the trajectories between two points in historical data, and to lay the foundation for trajectory prediction. Finally, the validity of the proposed method is verified on the real trajectory dataset.

Keywords: Uncertainties, kernel density estimation, modelling method, distribution characteristics.

1 Introduction

With the advent of lots of mobile equipments, the information about human location can be acquired. The trajectory historical data can be used to predict human mobility trajectory. The accurate and efficient prediction results could provide the required information and help to users, and improve the quality of location-oriented applications and services. Research on trajectory prediction include methods based on individual trajectory information entropy [1], trajectory frequent pattern mining [2-4], complex information-based behavior pattern mining [5,6], and hybrid methods [7]. However, human being mobility behaviors are extremely complicated, and there may exist many uncertain factors, such as objective environment and subjective intention, which will affect the accuracy of prediction. Although the mobility

EAI MOBIMEDIA 2020, August 27-28, Harbin, People's Republic of China

DOI 10.4108/eai.27-8-2020.2296731

behaviors are complex and changeable, in general, human behavior is more purposeful and the end point of a trajectory is relatively clear. So we can predict the end point of a trajectory.

Generally, the prediction method of the trajectory end point needs to match the trajectory to be predicted with the historical trajectories, and select the end point of the most similar trajectory to be the prediciton result. The matching methods of trajectories include calculating the similarity between trajectories based on distance [8], Markov model based on state transition probability [9-11], and Bayesian mixed model [12-14]. To judge the end point, almost all of them need to judge the matching degree between the trajectory to be predicted and the historical trajectory data. However, human mobility behaviors have strong uncertainties. A user's current trajectory may not be the same as any historical trajectory, but it may have the same start point and end point with a historical trajectory. It is difficult to calculate the matching degree between such trajectories containing uncertainties used the existing methods.

A trajectory is made up of several location information that users have passed. If a user's mobility trajectory is considered to be uncertain, the trajectories having the same start and end points constitute an uncertain trajectory dataset representing all passing routes between these two points. That is, an uncertain trajectory is a sample set containing all possible routes between two points. Then we could obtain the behavior distribution characteristics of human from all trajectories between two points in the uncertain dataset. This distribution characteristics of human behaviors are specifically embodied as the distribution characteristics of location information in the trajectory. If it is considered that the location points in the trajectories obey a certain density distribution, it is more likely to select the location having a higher probability density in a uncertain dataset, when analyzing the current trajectory of a user.

The kernel density estimation method could be used to construct a probability density function that is more according with the actual distribution without assuming the distribution type previously. Therefore, we will use its ideas, and propose a trajectory modeling method based on kernel density estimation. The constructed model represents the distribution characteristics of locations in the trajectories of two points, and lays a foundation for the end point prediction of a trajectory with uncertain information.

2 Uncertain Trajectory Modelling Method

The ideal trajectory information contain the start point, end point, necessary path information of objects, and have less noise or redundancies. This moving mode is simple and easy to calculate similarities between trajectories. However, in some real mobile scenarios, due to the complexity of people and society, human mobility behaviors have strong uncertainties, so that there may exist different trajectories when moving to the same target as the end point. For example, someone may choose a relatively long route to avoid traffic congestion, or choose a casual route for interests or special purposes. It is clear that on the one hand, it is difficult to accurately determine the details of the trajectories (e.g. the next accurate moving position), on the other hand, it is difficult to construct a model for a trajectory based on a single or simple hypothetical theoretical model. The existance of complexity and uncertainty make a challenge for trajectory modeling and analysis.

We do not simply ignore the uncertainty in the trajectory, but treat all the trajectories between two points in historical data as samples to build an uncertainty-oriented trajectory model between two points. So the uncertainties reflect all the possible routes from a start point to an end point, and can be reprensted as the distribution features of trajectories used probability density function. Based on the model describing the distribution of trajectories between points, and the individual's current movement route, we can analyze his most likely movement direction. The specific constructing trajectory model method is described as follows.

First, dataset D represents a historical trajectory set, in which each one records the location information of an individual collected at different time points in a mobility behavior, that is, each record consists of the object's location information series. Sequence composition. Then the trajectory set $D = \{T^1, T^2, \ldots, T^n\}$, where $T^i (1 \le i \le n)$ is the *i*-th trajectory sequence. It can be denoted as $T^i = \{u_1^i, u_2^i, \ldots, u_d^i\}(1 \le i \le n)$, i.e. an ordered set of positions collected at d time points of the *i*-th trajectory, in which $u_j^i = (x_j^i, y_j^i, t_j^i), 1 \le j \le d$ is the position information collected at the *j* time point, representing that two-dimensional position information (x_j^i, y_j^i) is collected at the sampling time t_j^i.

That is, the trajectory dataset consists of several trajectory sequences like T^i, which consists of position information (trajectory points) acquired at different time. The trajectory dataset can also be represented as $D = \{(u_1^1, u_2^1, \ldots), (u_1^2, u_2^2, \ldots), \ldots, (u_1^n, u_2^n, \ldots)\}$, then a location set U can be acquired: $U = \{u_1^1, u_2^1, \ldots, u_1^2, u_2^2, \ldots, u_1^n, u_2^n, \ldots\}$ through extracting location information from D. Actually, some locations in U may be the same or nearby with each other geographically. Similar to the setting of a bus stop, the locations close to a certain station belong to the station in the reachable range. If you aim at a destination close to a certain station, you can get off at that station in advance. In fact, locations with similar coordinates can be treated as the same in trajectory prediction and planning application. Similarly, we can obtain the set V containing trajectory points after processing the locations having similar coordinates in U as the same points. The integrated trajectory point set $V = \{v_1, v_2, \ldots, v_k\}$, apparently $|V| \le |U|$, is helpful for reconstructing historical trajectories and analyzing their characteristics between the same origin and destination.

Then we can also reconstruct the trajectory sequence $T^{i}{}'(1 \le i \le n)$ in D as $T^{i}{}' = \{v_1^i, v_2^i, \ldots, v_{d'}^i\}$, and there may exist several identical trajectory points in different trajectories. And the reconstructed trajectory dataset $D' = \{T^{1}{}', T^{2}{}', \ldots, T^{n}{}'\}$. We can extract trajectories from the trajectory sequences having the same origin and destination in D'. These trajectories may contain different trajectory points, which reflect different behaviors of different users. And this type of trajectories with the same start point s and end point e as well as passing through different trajectory point sets is called an uncertain trajectory UT between s and e. Its formal description is a quadruple $UT = (s, e, ns, ts)$, consisting of the start point s, the end point e, the trajectory point set ns, and the trajectory set ts it passing by, where $s, e \in V$, $ns \subset V$, ts is the set of all trajectories passing through s and e in D', which can be described as $ts_{s,e} = T_{s,e}^{1'} \cup T_{s,e}^{2'} \cdots \cup T_{s,e}^{n'}$.

Each data in D' represents an actual user's mobility trajectory sequence $T^{i}{}'$, in which there may exist at least one trajectory between any two points. For example, $T_{i,j}^{1'}$ is the reachable trajectory between v_i and v_j in $T^{1}{}'$. If we take $\forall v_i, v_j \in V$ as the start point and the end point of the uncertain trajectory respectively, and assume that each trajectory sequence in D' includes the routes between these two trajectory points v_i and v_j, we can obtain the uncertain trajectory $UT_{i,j} = \{(|v_i, v_j, ts_{i,j})|v_i, v_j \in V\}$ between v_i and v_j, where $ts_{i,j} = T_{i,j}^{1'} \cup T_{i,j}^{2'} \cup \cdots \cup T_{i,j}^{n'}$, representing all the routes that users are likely to choose when passing these two points in historical data.

Then, through analyzing the trajectory dataset D', we can get an uncertain trajectory dataset $UTD = \{UT_{i,j}\}$, where each element represents an uncertain trajectory sequence $UT_{i,j} = \{(|v_i, v_j, ts_{i,j})|v_i, v_j \in V, i \ne j\}$. It is clear that there is not only one reachable sub route between v_i and v_j in $ts_{i,j}$, but it covers all trajectories that users have passed by between two points in the collected historical data. Therefore, the routes between any two trajectory points could reflect the behaviors of different users in the mobility process, and could provide abundant information resources for the subsequent trajectory prediction.

Although it is difficult to analyze the purpose of individual when selecting their moving route based on the trajectory dataset, it is possible to analyze the distribution characteristics of mobility behaviors between two points through the trajectories contained in $ts_{i,j}$. The trajectory points passed by the trajectories in $ts_{i,j}$ are extremely complicated and uncertain, which may have randomness due to some objective environment or subjective intentions in the mobility behaviors. Therefore, the density distribution of the trajectories generally do not obey some simple hypothetical distribution, such as normal distribution, or power law distribution, etc. The kernel density estimation method can be used to obtain the probability density function consistent with the actual distribution characteristics without assuming data distribution previously. We will use the kernel density estimation method to obtain the distribution characteristics of the trajectory points in an uncertain trajectory.

The trajectories in the uncertain trajectory ts can be regarded as the samples representing the reachable paths between two points. When analyzing an uncertain trajectory, we will extract the samples about the reachable paths to get the uncertain trajectory between two points from the dataset, and use kernel density estimation to obtain the probability density function that can reflect the distribution characteristics of mobility behaviors in the uncertain trajectory. In addition, the trajectory points and their numbers may differ from each other in different trajectories, so it is difficult to directly analyze the distribution density of each uncertain trajectory. While all trajectories in ts are composed of trajectory points, so when analyzing the characteristics of behaviors between two points, we can refine the analysis objects to the trajectory points, then we can construct the probability density function to describe the distribution characteristics of trajectory points in the historical data.

As defined as above, $UT = (s,e,ns,ts)$ is the uncertain trajectory between the start point v_s and the end point v_e , $v_s, v_e \in V$. $ns_{s,e}$ is the set of trajectory points contained in all trajectories in ts , and $ns_{s,e} = \{v_1^{s,e}, v_2^{s,e}, \ldots, v_m^{s,e}\}, v_l^{s,e} \in V(1 \le l \le m)$, where m is the number of trajectory points. According to the kernel density estimation, we can obtain the probability density function of $ns_{s,e}$: $f^{s,e} = \frac{1}{mh_d}\sum_{l=1}^{m} K(\frac{x-v_l^{s,e}}{h^d})$, where x represents the d-dimensional independent variable. Since the probability density function is constructed for the trajectory points, the independent variable in this study is a 2-dimensional variable. h^d represents the optimal bandwidth with d dimensions. Then the density uncertainty trajectory between v_s and v_e is $UTf = (s,e,f)$.

In order to extract the information that is conducive to predicting and analyzing the user's trajectory behaviors, it is an important basis to extract the uncertain trajectories and construct their models between any two points. For an uncertain trajectory, the construction of its model includes not only the start point, end point of the trajectory, and multiple trajectories generated by the users between two points, but also the distribution characteristics of the trajectory points covered in these routes. That is, the uncertainty of a trajectory can be reflected as the multiple optional routes generated by different users, and its description can be a probability density function reflecting the distribution characteristics of the trajectory points contained in the possible routes. Actually, the model constructed no longer pays attention to the trajectory details difficult to determine, but pays more attention to the target of the trajectory and the possible trajectory patterns .

3 Experiments

We will verify the effectiveness of proposed method on GPS trajectory data Geolife collected by Microsoft Research Asia [15], a publicly available dataset commonly used in the field of trajectory research. This data is recorded by Map Life, including GPS coordinates of 182 users in 5 years, including latitude, longitude, altitude, and time. The dataset contains 17,621 trajectories, 24874410 location information, with a total distance of 1292951 kilometers and a total duration of 50176 hours.

In order to extract reasonable trajectory data, the start point and end point of a trajectory need to be reasonably determined, and the continuous position points between two points construct a reasonable trajectory. In addition to the original trajectory data in Geolife, any trajectory in the dataset may contain multiple sub-trajectories. Although these sub-trajectories do not belong to the actual complete trajectory in the original data, they are likely to be the basis for analyzing the mobility behaviors of other users. It is necessary to extract the information of these sub-trajectories and treat them as samples of uncertain trajectories. Considering that the users may break the original trajectory into several parts through some actions, such as resting and eating midway. These breaks can be seen as trjactory points and be used to extract sub-trajectories. Therefore, to obtain more abundant trajectory data, firstly we calculate the frequency of the retention time in the trajectories, and we can get frequency distribution of retention time generated by different users in different trajectories as shown in **Figure 1**.

Figure 1 shows that the retention time and its corresponding frequency tend to be linear in the coordinate system. The general statistical period in this trajectory dataset is 1 second or 5 seconds, and more than 5 seconds can be regarded as a break. When the retention time is longer than 10 seconds, its frequency decreases significantly. Therefore, the dwell points with more than 10 seconds are considered as a stagnation point. If the stagnation point belongs to a trajectory point in a trajectory, it is used to obtain the sub-trajectory of the original data. That

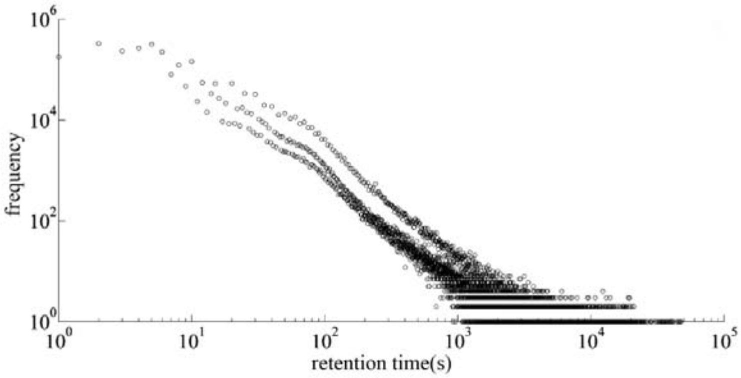

Fig.1. Frequency distribution of retention time

is, the The start and end points of an uncertain trajectory are defined according to the stagnation points. We obtain the trajectories passing through any two stagnation points from the original dataset, and treat them as the samples in an uncertain trajectory between these two stagnation points. Then we can construct its uncertain trajectory model: calculate the probability density function of the trajectory points contained in these trajectory samples, and analyze the distribution characteristics of the included trajectory points.

We will verify the effictiveness of the proposed modeling method through the prediction of target of a trajectory. Given a user's trajectory, we predict his destination, according to the current trajectory points passed by already, through matching with the trajectories extracted from the dataset. **Figure 2** depicts the results about prediction accuracy on the Geolife trajectory dataset. It can be seen that the prediction accuracy can reach 70% with 40% input data. As the size of the input trajectory data to be predicted continues to increase, the prediction accuracy continues to improve. Moreover, the variance of the prediction results is small, indicating that the proposed method not only has high prediction accuracy but also has good stability.

Conclusion

To reduce the impact of uncertainties on trajectory mining, this study uses the kernel density estimation method is to construct models for the uncertain trajectories and represent the distribution characteristics of their uncertainties through probability density functions. Kernel density estimation method could obtain more objective distribution characteristics of uncertain trajectories conforming to the actual distribution of data. The proposed method fully considers the uncertainties reflecting mobility behaviors and any possible location in the historical trajectories. Therefore, it can obtain more accurate modeling results and lay the foundation for trajectory prediction. Finally, experiments on the real dataset also verify the effectiveness and reliability of the proposed method.

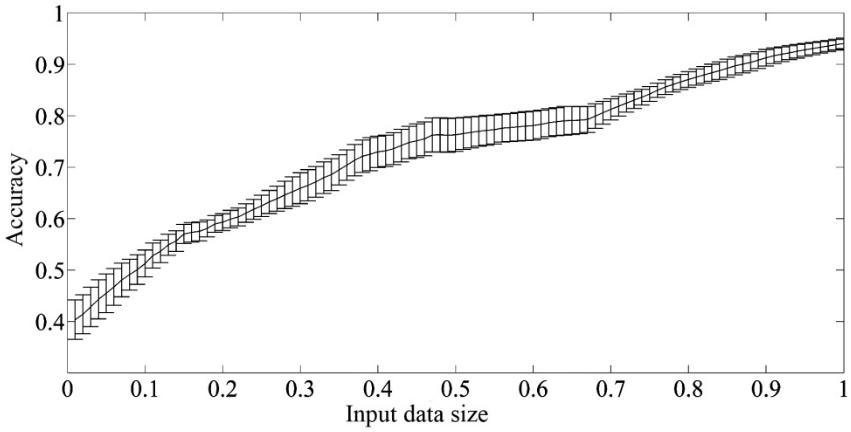

Fig.2. Prediction accuracy result

Acknowledgments. This work is sponsored by the Natural Science Foundation of Heilongjiang Province under grant number F2017015, and Training Program For Young Innovators in Heilongjiang General Institutes of Higher Education under grant number UNPYSCT-2017079.

References

[1]Song C, Qu Z, Blumm N, et al.: Limits of predictability in human mobility. Science,Vol. 327(No. 5968), pp. 1018-1021 (2010)

[2]Zheng Y, Zhang L, Xie X, et al.: Mining interesting locations and travel sequences from GPS trajectorie. Proceedings of the 18th international conference on World wide web. pp. 791-800(2009)

[3] Chen Z, Shen H T, Zhou X, et al.: Discovering popular routes from trajectories. 2011 IEEE 27th International Conference on Data Engineering. IEEE, pp. 900-911(2011)

[4]Jeung H, Liu Q, Shen H T, et al.: A hybrid prediction model for movingobjects. In: Data Engineering, 2008. ICDE 2008. IEEE 24thInternational Conference on. IEEE, pp. 70-79 (2008)

[5]Ying J J C, Lee W C, Tseng V S, et al.: Mining geographic-temporal-semantic patterns in trajectories for location prediction. ACM Transactions on Intelligent Systems and Technology (TIST), Vol. 5(No. 1), pp. 1-33 (2014)

[6]Gambs S, Killijian M O, del Prado Cortez M N, et al.: Next place prediction using mobility markov chains. Proceedings of the First Workshop on Measurement, Privacy, and Mobility.pp. 1-6(2012)

[7]Qiao Y, Si Z, Zhang Y, et al.: A hybrid Markov-based model for human mobility prediction. Neurocomputing,Vol. 278, pp. 99-109(2018)

[8]Toqué F, Côme E, El Mahrsi M K, et al.: Forecasting dynamic public transport origin-destination matrices with long-short term memory recurrent neural networks. 2016 IEEE 19th international conference on intelligent transportation systems (ITSC). IEEE, pp. 1071-1076(2016)

[9]Ashbrook D, Starner T.: Using GPS to learn significant locations andpredict movement across multiple users. Personal and Ubiquitouscomputing,Vol. 7(No. 5), pp. 275-286(2003)

[10]Mathew W, Raposo R, Martins B, et al.: Predicting future locations with hidden Markov models. Proceedings of the 2012 ACM conference on ubiquitous computing.pp.911-918(2012)

[11]Alvarez-Garcia J A, Ortega J A, Gonzalez-Abril L, et al.: Trip destination prediction based on past GPS log using a Hidden Markov Model. Expert Systems with Applications,Vol. 37(No. 12), pp. 8166-8171(2010)

[12]Endo Y, Nishida K, Toda H, et al.: Predicting destinations from partial trajectories using recurrent neural network. Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Cham, pp.160-172(2017)

[13]Ziebart B D, Maas A L, Dey A K, et al.: Navigate like a cabbie:Probabilistic reasoning from observed context-aware behavior. In: Proceedings of the 10th international conference on Ubiquitous computing. ACM, pp. 322-331(2008)

[14]Al-Sultan S, Al-Bayatti A H, Zedan H. Context-aware driver behavior detection system in intelligent transportation systems. IEEE transactions on vehicular technology, Vol. 62(No. 9), pp.4264-4275(2013)

[15]Jing Yuan, Yu Zheng, Xing Xie, and Guangzhong Sun, et al.: Driving with knowledge from the physical world. In: In The 17th ACM SIGKDD International conference on Knowledge Discovery and Data mining, KDD'11, New York, NY, USA, 2011. ACM.USA (2011)