ATLANTIS PRESS

Advances in Computer Science Research, volume 80

International Conference on Computer Science, Electronics and Communication Engineering (CSECE 2018)

Identifying Route Preferences over Origin-Destination Using Cellular Network Data

Zhichao Guo and Tongyu Zhu*

State Key Laboratory of Software Development Environment, Beihang University

*Corresponding author

***Abstract*—Current research on studying people's routing behavior focus on how to provide the minimum cost routes while ignore users' preference. Therefore, to analyze user's routing preference, identifying routes they actually choose is a crucial task. Cellular network data contains sufficient spatio-temporal information, which is widely used for trajectory analysis nowadays. However, it is a big challenge to extract precise trajectory from the cellular network data due to its low positioning accuracy. Compared to Density-Based Spatial Clustering of Applications with Noise algorithm, we present a Spatio-Temporal Density Clustering algorithm considering the timing sequence of the points to promote the precision of the trajectories, which aims to filter users' most probable routes with the map matching algorithm. Our approach could find out those most probable routes and the probability of each route. Finally, we experimented with real data. The results show that our approach is efficient for both extracting the probable trajectories and identifying multi-routes that users would prefer to route.**

Keywords—cellular network data; spatio-temporal data; trajectory data mining; origin-destination

I. INTRODUCTION

Analyzing millions of urban inhabitants' mobility pattern is an important research field of urban computing. Many research on urban computing, such as transportation, urban planning and energy, are based on the analysis of people's moving regulations. Extracting people's trajectories and knowing how they move between key locations are fundamental to enhance our understanding of people's daily mobility pattern. Analysis of routing behavior could promote the performance of route recommender systems, help improve transportation infrastructure and reduce traffic congestion. Traditional route recommender systems are based on the assumption that individuals would choose the route that minimizes a cost, usually distance or time. However, some research has shown that many users choose some other routes, rather than the minimum cost path [1]. In fact, a user would choose different route with different purpose even with the same origination and destination. This shows results obtained by traditional optimal road selection theories are not credible when applied to route recommender systems which would try to learn users' preferences, and provide the most probable routes based on the analysis of users' real travel data. However, identifying the routes people really choose and analyzing their routing preference is not a easy task since it's hard to collect those privacy data from users since they are not willing to provide.

Some kinds of data have been used for Origin-Destination analysis, like volunteers' GPS data, probe vehicle data and bus IC card data. But all of them have a fatal limitation: They all only provide the sample dataset which is too small to represent all people's preferences. Fortunately, the widespread use of mobile phone made it possible to capture everyone's trace [2]. Cellular network data which collected by telecom operators provides rich spatio-temporal information about all the phones accessed to the cellular network. Whereas cellular network data has two defects. One defect is its bad positional accuracy. The location of mobile phone is estimated of the cell tower's signal coverage, thereby the accuracy is much lower than GPS system. According to the statistical data from CMCC(China Mobile Communication Corp), the positioning accuracy of mobile phone is about 50-100 meters. The other one is its poor sampling rate. In most cellular infrastructures today, mobile phones leave a record of their connected cell tower only when some specific event occurs. These two characteristics of cellular network data present a great challenge to extract individual's fine-grained trajectory.

However, we can still extract those probable routes from cellular network data since routes can be regulated as points in a trajectory which can be matched to several real paths with map matching method. In this paper, we extract individual trajectory from cellular network data and acquire different routes over Origin-Destination matching to real paths, such as roads, subway. We define user's daily trajectory with a sequence of stay points and pass-by points. We propose a modified density-based spatial clustering algorithm, named Spatio-Temporal Density Clustering (STDC), to identify stay points. Compared to traditional spatial density clustering (i.e. DBSCAN), the expanding of cluster of STDC not only considers the ε-neighborhood in space, but also considers the adjacency in timing sequence. Therefore, STDC could distinguish clusters in the same location with different times. Then, an approach mining different routes over given Origin-Destination is presented. This approach uses Dynamic Time Warping algorithm to measure the similarity of two trajectories, and uses OPTICS clustering and map matching algorithms to find out the routes. Finally, experiments were conducted with real cellular network data in the city of Beijing to validate our methods. The results show that both STDC algorithm and routes mining method are effective.

II. RELATED WORK

Methods of extracting trajectories from GPS are widely developed nowadays. On the contrary, how to use cellular network data is quite a challenge due to its low positioning accuracy and unstable sampling rate. For this reason, a large number of institutions have made beneficial explorations.

Leontiadis developed an algorithm to parse the continuous sector observations and identify the stationary and mobility segments [3]. This algorithm mainly focuses on the preprocessing of sectors, additional information are needed. Isaacman proposed a technique based on clustering and regression for analyzing anonymized cellular network data to identify generally important locations, such as home and work [4,5]. However, this method didn't generate user's daily trajectory. Although many researchers are trying to extract trajectory from cellular network data, there is still a great challenge to find stay points with a high spatial and temporal accuracy. Therefore, we present a modified density-based spatial clustering algorithm named Spatio-Temporal Density Clustering (STDC) to find stay points, which considers the temporality of the sequence and could distinguish the multiple clusters in the same spatial location with different time.

Recently, some institutes focus on Origin-Destination analysis using users' trajectories. Iqbal proposed a methodology to develop OD matrices using cellular network data [6]. They combined the cellular network data and traffic counts data and generated a node-to-node transient OD matrix and they calculated the traffic flow in origin nodes and destination nodes. Alexander presented a method to produce OD trips by purpose and time of day [7]. They calculated the distribution of trip length and flows between home and work locations using mobile phone data. However, their research is mainly focusing on the origin and destination, didn't pay attention to people's moving pattern and routing behavior between origin and destination. Lima used GPS traces generated by 526 private cars to explore their routing behavior [8]. They extracted significant locations, firstly. Then they deployed a clustering algorithm to detect routes. However, the stability of this method relies on the accuracy of GPS data and the clustering algorithm is sensitive to the setting of ε. Therefore, we adopt a clustering algorithm that is not sensitive to the position accuracy in order to adapt for cellular network data.

III. DATASET DESCRIPTION AND PRE-PROCESSING

In this paper, we use a dataset consisting of anonymous cellular network data collected by telecom operator during a month in the area of Beijing. General information of the dataset is summarized in Table 1. Each record contains an anonymous user ID, timestamp, longitude, latitude, cell type, event ID, etc.

TABLE I. CELLULAR NETWORK DATASET INFORMATION

Dataset information	value
Number of calls	750.5 million per day
Number of users	14.9 million
Average update cycle	21.7 min
Population of beijing	21.15 million
Area of data coverage	16410.54 km^2

A. User Filtration

Cellular network data are generated when a phone connects to the cellular network including making or receiving calls, sending or receiving text messages, crossing Location Area Codes (LACs), routinely collected by the network due to stale data that exceeds a pre-configured duration.

Therefore, due to the difference of users' uptime and activeness, the frequency of different users generating records varies greatly. Some users generate just a few records a day, which is not enough for extracting daily trajectory. After analysis, we filtered out about 70% users who generate records at least every 3 hours for the following analysis.

IV. INDIVIDUAL TRAJECTORY EXTRACTING

Identifying stay points from raw records is the core task of extracting trajectory. A stay point is generated from a series of raw records which are clustered in space and continuous in time. Density-based spatial clustering, like DBSCAN, is one of the best ways to identify clusters from the sequence of raw records. However, DBSCAN doesn't consider the temporality of the sequence, therefore it couldn't distinguish the multiple clusters in the same spatial location with different time. As shown in Figure 1.

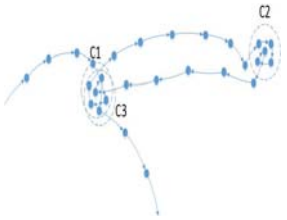

FIGURE I. EXAMPLE OF MULTIPLE CLUSTERS IN THE SAME LOCATION AT DIFFERENT TIME

We present a modified density-based spatial clustering algorithm named Spatio-Temporal Density Clustering (STDC). STDC considers the temporality, and optimizes for the low positioning accuracy and unstable sampling rate of the cellular network data. STDC inherits and modifies the definition of DBSCAN. In addition, it changes the clustering process and adds an input parameter compared to DBSCAN.

A. Definitions for Trajectory

- **Definition 1.** Raw record is pre-processed cellular network data, ignore the field of no use. We represent a raw record as:

$$\text{RawRecord: R(id, t, lon, lat)} \quad (1)$$

Which means a user appeared in the location of (lon,lat) at the time of t.

- **Definition 2.** Stay point is a part of trajectory, which indicates a user keep in a certain space (i.e. office building) for a certain time interval. A stay point SP is generated by a raw record cluster RC:

$$\text{StayPoint: SP}\left(RC_{(lon,lat)}, RC_{(ts,te)}\right) \quad (2)$$

Where $RC_{(lon,lat)}$ represents the geometric center of RC and $RC_{(ts,te)}$ represents the starting and ending time of RC:

$$RC_{(lon,lat)} = \left(\frac{1}{n}\sum_{i=1}^{n} R_{(i,lon)}, \frac{1}{n}\sum_{i=1}^{n} R_{(i,lat)}\right), R \in RC \quad (3)$$

$$RC_{(ts,te)} = \left(\min\left(R_{i,t}\right), \max\left(R_{i,t}\right)\right), \mathrm{R} \in \mathrm{RC} \quad (4)$$

- **Definition 3.** Pass-by point is a part of trajectory, which indicates a passed place when a user move from a stay point to another one. A pass-by point is generated by a raw record which unable to gather into a cluster.

$$\text{PassbyPoint: } \mathrm{PP}(R_{(lon,lat)}, R_t) \quad (5)$$

- **Definition 4.** Trajectory is defined as a sequence of stay points and pass-by points to represent a user's trace during one day:

$$\text{Trajectory: } \mathrm{TR}(SP_1, SP_2, SP_3, \ldots, SP_n) \quad (6)$$

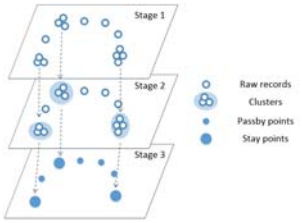

FIGURE II. DEFINITIONS FOR TRAJECTORY

B. *Definitions for STDC*

- **Definition 5. ε**-neighborhood: The **ε** -neighborhood of a point p is the region with the radius of **ε** by the center of p.
- **Definition 6.** Directly timing density reachable: A point q is directly timing density reachable from p if q in the **ε** -neighborhood of p and q is the following point of p in the sequence of time.
- **Definition 7.** Timing density reachable: A point q is timing density reachable from p if there is a path p_1, ..., p_n with p_1 =p and p_n =q, where each p_{i+1} is directly timing density reachable from p_i.
- **Definition 8.** Core point: A point p is a core point if the time span of all the points timing density reachable from it (including p) is more than MinTimeSpan.

C. *STDC Approach*

STDC algorithm works as follows to visit all points:

1) In order to find clusters, STDC firstly marks all points in the dataset for unvisited, and sorts the points by timestamp.

2) STDC selects the first unvisited point p from the sorted sequence and marks it for visited. If p is a core point, STDC will create a new cluster C and add p into C. Then it will create a set N for points reachable from p.

3) For each unvisited q in N, if q is a core point, then add its' all reachable points to N; if q does not belong to any cluster, then add q to C. Until all points in N are visited output C. Otherwise output p as a pass-by point.

4) Repeat the above steps until all points are visited.

V. MULTI-ROUTES MINING

Routes mining is divided into three steps:

1) Quantitative description of the difference between two trajectories is critical. We propose a method based on dynamic time warping (DTW) algorithm to calculate the distance between trajectories.

2) We use clustering approach to distribute them into different clusters. Because the number of clusters cannot be known in advance, so we use density-based clustering, OPTICS, rather than the algorithms which need to specify cluster numbers, such as K-means.

3) We implement a map matching process to match the trajectory cluster with the road.

A. *Measure Distance between Trips*

Dynamic Time Warping (DTW) uses a dynamic programming approach to align the time series and a specific word template so that some distance measure is minimized. The pattern detection task involves searching two time series P with the length of n, and Q with the length of m.

$$\mathrm{P} = p_1, p_2, p_3, \ldots, p_n \quad (7)$$

$$\mathrm{Q} = q_1, q_2, q_3, \ldots, q_n \quad (8)$$

The first step is to define a n-by-m distance matrix, where each grid point (i,j) corresponds to an alignment between elements p_i and q_i.

$$\boldsymbol{D_{Matrix}} = \begin{bmatrix} \boldsymbol{d(p_1, q_1)} & \cdots & \boldsymbol{d(p_1, q_m)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{d(p_n, q_1)} & \cdots & \boldsymbol{d(p_n, q_m)} \end{bmatrix} \quad (9)$$

Where $d(p_i,q_i)$ is the distance between two time series. The definition of distance is the foundation of clustering.

The next step is to define a sequence of matrix element to represent the warping path.

$$\mathrm{W} = \mathrm{w}_1, \mathrm{w}_2, \mathrm{w}_3, \ldots, \mathrm{w}_k \quad (10)$$

According to the analysis of the distance matrix, warping path may have more than one solution. In this paper, we only concern the minimum one for simplicity. In logically, the maximum of similarity between the two time series as the criterion of the similarity search.

$$\mathrm{DTW(P, Q)} = \min\left(\frac{1}{K}\sum_{k=1}^{K} w_k\right) \quad (11)$$

We use the average of w_k rather than the maximum value because the maximum value may affected by the noisy data which would lead to a wrong result.

B. *Trajectories Clustering*

Trajectories clustering is dividing trajectories into different clusters according to the similarity measure. The objects in the same cluster have a higher similarity, while the objects in different clusters have greater differences. Meanwhile,

clustering analysis for trajectories can also be used as a pre-processing step for other algorithms, like classification and pattern extracting.

In this paper, we use OPTICS algorithm for trajectories clustering, which based on the distance calculated by DTW. We would not use DBSCAN here because DBSCAN is sensitive to the value of ε and minPts, meanwhile the similarity distance of trajectories is sensitive to the distance between origin and destination. Although ε and minPts are also needed in OPTICS, they only play an assistant role in this algorithm.

C. Map Matching

Map matching is the process of matching trajectory with the road in a digital map, which could find the user's real route in the road network. After that, we analyze the means of transportation the user actually choose in this route. The research for map matching algorithm has mature results in the field of floating car data processing. We adopt a heuristic map matching algorithm in this research [9].

VI. Experiment and Result

We implement three experiments to validate the effectiveness of our methods. Firstly, we choose the suitable parameters for STDC algorithm. Secondly, we selected user's trajectory to compare their raw records and stay points extracted by STDC algorithm. Finally, we implement the routes mining process for several Origin-Destination pairs.

A. STDC Parameters Setting

In order to select appropriate parameter of ε, we draw a scatter diagram for the distribution of distances between adjacent records. As shown in Figure 3, there is a peak in the interval of (500, 600). This indicates that most adjacent stay points have a distance about 500m, so the ε we choose must bigger than it. Hence, we set 800 meters as the value of ε and 30 minutes as the value of MinTimeSpan in this paper.

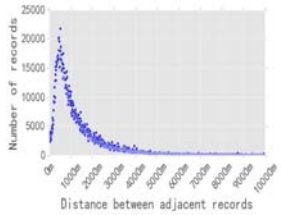

FIGURE III. DISTRIBUTION OF DISTANCE BETWEEN ADJACENT RECORDS

B. Stay Point and Trajectory

This experiment verified the effectiveness of the STDC algorithm through comparing user's raw records and stay points. We selected a user, as shown in Figure 4, he/she has three stay points in his/her one-day trajectory. The first stay point (01:22-06:54) and the third stay point (19:16-23:55) at the same location. The results shown that STDC algorithm could correctly distinguish the stay points at the same location with different times.

FIGURE IV. EXTRACTED USER'S STAY POINTS

C. Multi-routes over Origin-Destination

In this section, we choose 2 typical commuting OD pairs, OD1: Tiantongyuan-Guomao and OD2: Tongzhou-Zhongguancun. Firstly, we extract all trajectories over the Origin-Destination pair and use our method to cluster them. We compared the clustering results got by OPTICS and DBSCAN, we set ε= 800 meters and minPts = 10 for both OPTICS and DBSCAN, the results are shown in Figure 5. For both OD1 and OD2, OPTICS could effectively distribute the trajectories into different clusters and discard the noisy data which could not be used. On the contrary, the DBSCAN's results are not obvious. It is because DBSCAN is sensitive to the parameters setting so that many trajectories are improperly divided into one cluster and many noisy data couldn't be filtered. It shows that OPTICS performs better in this scenario.

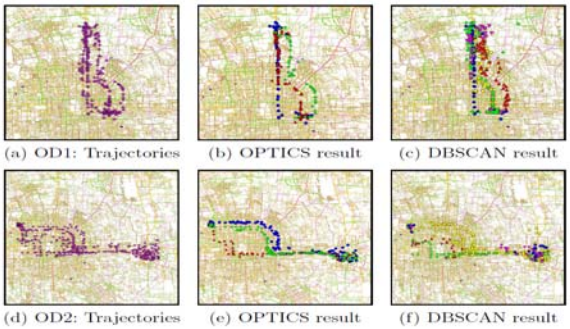

FIGURE V. TRAJECTORIES CLUSTERING RESULTS

Next, we matched the trajectory clusters with the digital map(in this paper, we use Baidu map). Combined with the subway data and bus data, each cluster is matched with a route, as shown in Figure 6. For OD1, the route 1 matched with a bus route, the route 2 matched with a subway route and the route 3 matched with a self-driving route. The ratio of the users who choose these three kinds of routes is about 5:3:2. For OD2, the route 1 and route 2 matched with subway routes and the route 3 matched with a bus route or self-driving route. The corresponding ratio is about 2:4:4. The map matching results show that that the trajectories we extracted are appropriate and the routes we identified are reliable.

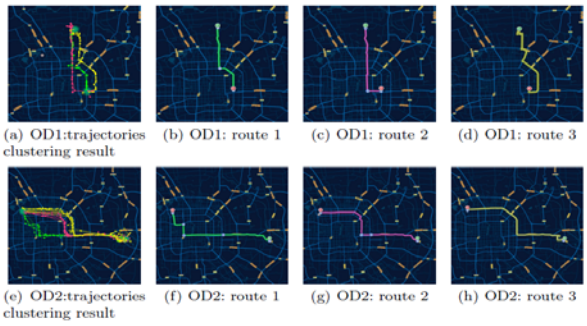

FIGURE VI. MAP MATCHING RESULTS

VII. CONCLUSIONS

How to analyze user's routing preference using cellular network data with low positioning accuracy is a challenge task. Compare with other approaches, we try to present those most probable routes with their probability between Origin-Destination through a novel algorithm named as STDC that extracts user's trajectory from cellular network data and mines multi-routes between Origin-Destination. First of all, we pre-processed the raw data to filter out noisy data. Then we defined the user's daily trajectory and present a new algorithm to identify stay points. This algorithm considered the temporality of points sequence. Then a new approach was presented for mining routes between Origin-Destination. This method used DTW algorithm to measure the similarity between trajectories and used OPTICS divided them into different clusters. Each cluster matched with a route between this Origin-Destination. Finally, we conducted a series experiments to validate our method is effective and practicable. The experimental results show that cellular network data could be used for identifying the different routes between Origin-Destination and count how many people choose the routes.

Our approach and experimental results could be used for urban traffic planning, and even avoid traffic congestion since we know users preferences and change traffic condition in advance. Our future work will focus on trajectory filling techniques, which uses user's a number of days of incomplete trajectories to generate a complete trajectory. This could help us to improve the precision of the method.

ACKNOWLEDGMENT

This research is supported by the National High Technology Research and Development Program of China (863 Program) No.2015AA124103.

REFERENCES

[1] Lima, A., Stanojevic, R., Papagiannaki, D., Rodriguez, P., & González, M. C. (2016). Understanding individual routing behaviour. Journal of the Royal Society Interface, 13(116), 20160021.

[2] Lane, N. D., Miluzzo, E., Lu, H., & Peebles, D. (2010). A survey of mobile phone sensing. IEEE Communications Magazine, 48(9), 140-150.

[3] Leontiadis, I., Lima, A., Kwak, H., Stanojevic, R., Wetherall, D., & Papagiannaki, K. (2014). From Cells to Streets: Estimating Mobile Paths with Cellular-Side Data. ACM International on Conference on Emerging NETWORKING Experiments and Technologies (pp.121-132). ACM.

[4] Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., & Rowland, J., et al. (2011). Identifying Important Places in People's Lives from Cellular Network Data. International Conference on Pervasive Computing (Vol.6696, pp.133-151). Springer Berlin Heidelberg.

[5] Isaacman, S., Becker, R., Martonosi, M., Rowland, J., Varshavsky, A., & Willinger, W. (2012). Human mobility modeling at metropolitan scales. Proceedings of the 10th international conference on Mobile systems, applications, and services (pp.239-252). ACM.

[6] Iqbal, M. S., Choudhury, C. F., Wang, P., & González, M. C. (2014). Development of origin–destination matrices using mobile phone call data. Transportation Research Part C Emerging Technologies, 40(1), 63-74.

[7] Alexander, L., Jiang, S., Murga, M., & González, M. C. (2015). Origin–destination trips by purpose and time of day inferred from mobile phone data. Transportation Research Part C Emerging Technologies, 58, 240-250.

[8] Lima, A., Stanojevic, R., Papagiannaki, D., Rodriguez, P., & González, M. C. (2016). Understanding individual routing behaviour. Journal of the Royal Society Interface, 13(116), 20160021.

[9] Wu, D., Zhu, T., Lv, W., & Gao, X. (2007). A Heuristic Map-Matching Algorithm by Using Vector-Based Recognition. International Multi-conference on Computing in the Global Information Technology (pp.18). IEEE Computer Society.