Urban Trace Utilizing Mobile Sequence

Yukun Ma(✉), Bin Xu, and Qi Li

Department of Computer Science and Technology, Tsinghua University, Beijing, China
yukunma@aliyun.com, xubin@tinghua.edu.cn, zhongguoliqi@163.com

Abstract. Urban trace mining and modeling of large scale active mobile phone holders is of great significance as it can effectively reduce traffic jam and hit-and-run incidents as well as other urban problems. Traditionally, people monitors traffic flows by GPS devices installed on buses or cars and traces hit-and-run incidents by road cameras, which turns out to be inefficient in that the data are sparse to represent a massive of people traces. In this paper, we propose a novel approach called mobile sequence to describe urban roads and moving object trajectories. Firstly, we use real urban roads information and base stations location information to generate urban roads mobile sequence. Then we extract people's mobile sequence from massive mobile network logs. Finally, we match roads' mobile sequence with people's mobile sequence to obtain human traces. As a validation, we use application installed on mobile phone to gather some real urban roads and base stations' information and then calculate their real mobile sequences to match the theoretical mobile sequences and gain 94.8% covering rate. Also, we use people's mobile sequences extracted from mobile network logs with theoretical mobile sequences and gain 90.6% matching accuracy.

Keywords: Urban trace · Mobile sequence · Urban real roads · Mobile network logs

1 Introduction

With wider and deeper urbanization of the world, big cities emerge in a large number, which have engendered many great issues involving environmental pollution, energy wasting and many transportation problems. Facing with the kinds of great challenges, governments and researchers proposed a myriad of creative solutions. Among all of the novel tackles, methodologies based on big data are repeatedly mentioned because of the advent of the era of big data. Additionally, more conceptions and assumptions to solve modernization issues are put forward, such as smart cities [3], urban computing [20], etc.

When people use intelligent phones to access the Internet, user behavior data are recorded in logs by operator. If we know a majority of people's traces in advance, then some urban issues, such as traffic congestion and hit-and-run incidents, can be eased and reduced efficiently. At present there are mainly three

B. Zou et al. (Eds.): ICPCSEE 2017, Part II, CCIS 728, pp. 241–255, 2017.
DOI: 10.1007/978-981-10-6388-6_20

kinds of methods to detect a road traffic jam. The first method is based on GPS of the mobile terminal. When related applications in mobile holders are started, they will collect GPS location information to upload to the server and the server will compute the terminal speed so that we can judge if there is a traffic congestion. The second method is to cooperate with special industries, such as bus companies, taxi companies. In the industries, cars will be installed GPS sensor, then the company can collect a huge of real-time GPS data. The data will be combined with the historical big data at the same period to compute the speed of the car in this road section. The last method is based on the detecting roadside cameras or speed radar equipment so as to detect and monitor the number and the rough speed of vehicles and to estimate the trafficability of a road. The forementioned methods are mainly based on GPS data or depend on some particular devices to collect vehicles' precise location information, which is not so convenient and economical. In contrast, we propose a novel method to detect the traffic anomalies based on mobile big data from mobile operator. The main challenge, when we apply mobile network data to modeling and analysis, is that we need extract spatio-temporal data as accurately as possible from people's Internet behavior traces. One reason is that the granularity of the base station coverage ranges from dozens of meters to hundreds of meters, which is determined by the stations' density in that area. If we can compute the average speed of the mobile devices, then we can determine a person's transportation ways, such as walking or by car, and we can also judge if the person is on the highway. In this paper, we propose an approach only based on people's mobile phones to solve the challenge. In our approach, urban roads and people's traces are represented by mobile sequence. The approach's effectiveness was demonstrated by real urban roads information collected by specialized applications installed in mobile phones. When we make a strict matching between the theoretical mobile sequence and the practical mobile sequence, we obtain about 74% correctly matching rate. However, almost 95% practical mobile sequences are covering by theoretical value, which means we just change the matching strategy then we can mine almost all the mobile users' trace based on the approach. In summary, the contributions of our work are as follows:

- We extract all of the users' daily traces from their mobile network logs and solved traces mining problem more conveniently and massively;
- We proposed a novel approach called mobile sequence to depict urban roads and people's urban traces;
- We provide a more convenient and efficient method to solve traffic issues.

The rest of the paper is organized as follows. In Sect. 2, we describe related work on urban trace data mining based on network data generated from mobile phones. Section 3 describes some preliminaries and some definitions. The novel approach we proposed to describe moving object traces in this paper will be demonstrated in Sect. 4. Section 5 shows the mobile sequences matching procedures and gives detail analysis of the results. Finally, we conclude this paper in Sect. 6 and also some future work will be illustrated in this section.

2 Related Work

Until now, lots of research on the city computing of urban trace mining remains emerge in endlessly [19]. On one hand it benefits from the convenience of the advance in location-acquisition technologies, which results in a myriad traces data including people, cars or other objects' mobility. Actually, the unprecedented traces data foster a broad range of novel thought patterns on smart cities, especially in location-based social networks [21] and urban computing [20]. On the other hand, in smart cities and urban computing, urban trace data mining has become an increasingly hot and productive research orientation.

Generally speaking, urban traces, after some data preprocessing, may lead us to discovery some predictable patterns or trends of urban dynamics [2,6,7,13,14]. Though a myriad of data categories in urban trace, a user carrying a mobile phone unintentionally generates traces recorded by a sequence of cell tower IDs. Thus, a phone naturally becomes a better sensor of a person's daily whereabouts [9]. If a person keeps open the phone flow everywhere, his Internet behavior traces are tracked by network logs. Even he just keeps his phone power on, his locations are passively tracked because his phone must communicate with the cellular network. All of the traces data are mined from big data and modeled to infer people's behavior [4]. Also, mobility trace data are applied to the population movement and patterns minning [12], but some universal features should be taken into consideration. Some researches pay more attention on the presence of preferential locations of one person's trace [13,14]. As for urban trace model, modeling is even more fine-grained over time. In 2002, White and Wells [15] considered moving objects' start points and end ports and proposed OD (Origin-Destination) Matrix to describe moving objects' traces to avoid more accurate traces problem only based on mobile data. According to the OD matrix, Iqbal et al. [8] combines the base stations with the traffic flows to monitor people movement flow. In 2007, POI (Point of Interest) [5] was proposed to discovery a part of people's similar interests. Because of different cover areas of the base stations, much more fine-grained traces are difficult to gather only from cellular network and mobile data. In 2015, Chao et al. [16,17] explored the location relationship between three base stations and the person who is covered simultaneously by three stations. They evaluate the person's position by triangulation method and the error distance reaches hundreds meters. Some researcher [10,18] take the spatio-temporal data mining as the frequent pattern problems to solve the trace mining problem.

In academia, two main trajectory technology methods of wireless location are network-based and handset-based. Network-based technology refers to analyze people's network behavior data when they are moving in a city and keep accessing the Internet. One problem is that people's Internet duration is limited in the past but now the data are intensive and dent because of the real-time applications' prevalence, such as Wechat, QQ. People usually keep their real-time applications on even they are driving, walking or eating. Thus, the task of depicting a user's trace from the network data is enough because high accuracy trace of a person may violate privacy, which is illegal. Based on the fact, handset-based technology, on one hand, we can use sensors in the mobile, such as GPS receiver to collect peoples' locations and have a relatively higher accuracy. On the other hand, almost all of the people are not willing to expose their location data unconditionally to other people or institutions. That means we can gather people's network data instead of their accurate location data, which is unobtrusive and continual.

Different from the methods above, in this paper, we will introduce a novel representation method based on a series of mobile base stations. A person moves from a base station to another base station and his position changeover will be covered by a new network called Thiessen polygon [1,11] network, composed of many Thiessen polygon instead of simple cellular network. The replacement reason will be illustrated in Sect. 3. We use this method to represent every moving objects' traces, and to match new traces so as to tackle the traffic congestion and other track problems.

3 Preliminaries and Definitions

Cellular network or mobile network is a mobile communication architecture, which is divided into analog cellular network and digital cellular network. In the real world, standard cellular network is an idealized model, which is influenced by some other geological and climatological conditions, such as buildings, humidity and temperatures. Therefore, the coverage of the base stations is not a standard cellular network and the standard network is unfeasible to indicate the coverage of one base station. An enlightment is from meteorology, then we alternatively choose Thiessen polygon [1,11] to divide the service coverage of all the base stations, which works fine practically. Figure 1 shows the real-world base stations network divided by Thiessen polygon.

In this paper, we probe into urban traces based on mobile network logs, which lead us give some definitions about human spatial-temporal trace and our

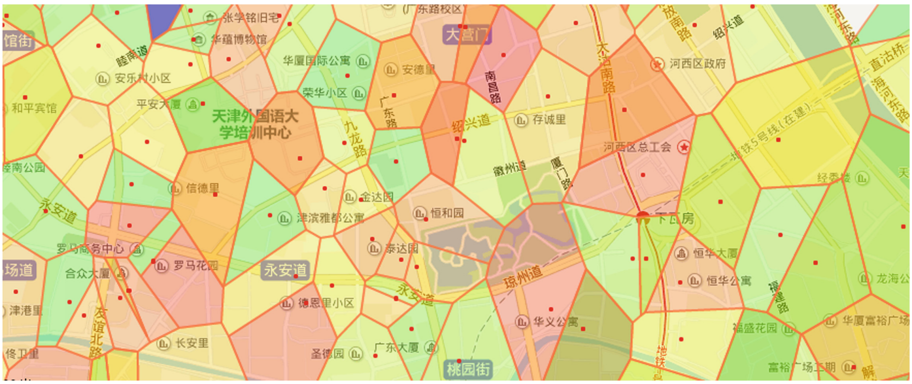

Fig. 1. Actual service coverage divided by Thiessen polygon in cellular network. (Color figure online)

new method called mobile sequence. As a rule, mobile base stations are always built along the roads. The whole road is always covered by some base stations, so we define RMS (road mobile sequence) to represent all the urban roads based on Thiessen polygon. Besides, after we extract human traces from network logs, we need define PMS (people mobile sequence) to stand for human traces in our mining algorithm. So the definitions are as follows:

Definition 1: ***point sequence of spatial-temporal trace.*** *It refers to a tuple* $T_p = \{p_0, p_1, \cdots, p_i, \cdots, p_n\}$, *where* $p_i = \{x_i, y_i, t_i\}$ *is an coordinate of the space.* x_i *is longitude,* y_i *is latitude and* t_i *is the time.*

Definition 2: ***human trace.*** *It refers to a tuple* $T_e = \{p_0, p_1, \cdots, p_j, \cdots, p_m\}$, *where* p_j *is the same as definition 1.*

Definition 3: ***mobile sequence.*** *It refers to a trace-around base stations' sequence which are generated by the base stations that cover the trace. Mobile sequence (MS) is a sorted set contained a series of base stations. So we define* $MS = \{BS_1, BS_2, \cdots, BS_i, \cdots, BS_n\}$, *where* BS_i *is the i-th base station in the sequence.*

Definition 4: ***RMS.*** *It refers to a mobile sequence generated from urban real roads based on Thiessen polygon algorithm.*

Definition 5: ***PMS.*** *It refers to a mobile sequence generated from a human trace extracted from network logs.*

4 Method

In this section, we will generally describe our method. Firstly, we use urban real roads network information and mobile base stations information to generate RMS (see definition RMS) based on Thiessen polygon algorithm. Secondly, we give the approach to generate PMS based on human moving sequences extracted from mobile network logs. Finally, we match PMS with RMS and obtain all the urban human traces.

4.1 Generating Road Representation Base on Mobile Sequence (RMS)

In this subsection we provide a simple RMS generating algorithm in Algorithm 1. After generating RMS, how does the forementioned RMS represent the urban road? Now we give the details of the representative method according to the Fig. 2. In Fig. 2, *road_A*, *road_B*, *road_C* and *road_D* are four roads and the cross of every two roads are assigned by the two road labels. For instance, the cross of *road_A* and *road_B* is labeled as AB and the rest can be done in the same manner. Now every road can be represented by the labels. (For example, $[A1, AC, AD, A2]$ represents *road_A*). In Fig. 2, only the base station correlated with road network are labeled but the irrelevant ones are not.

Algorithm 1. Generate RMS Algorithm

Require: Urban Real Roads Set: RS=$\{rs_1,rs_2,\ldots,rs_n\}$
Require: Base Stations Set: BS=$\{bs_1,bs_2,\ldots,bs_m\}$
$TN \Leftarrow []$;$CC \Leftarrow []$;$TP \Leftarrow []$;$RMS \Leftarrow \{\}$;
for bs_i in BS **do**
  Construct Delaunay triangulation network;
end for
for bs_i in BS **do**
  Find all the triangles numbers adjacent to bs_i save in TN;
  Sort every triangle adjacent to bs_i clockwise or anticlockwise;
end for
for tn_i in TN **do**
  Compute all the centers of triangles' circumscribed circle save in CC;
end for
for cc_i in CC **do**
  Connect all the centers of triangles' circumscribed circle according to every neighbouring triangle to generate Thiessen polygons in TP;
end for
for rs_i in RS **do**
  for tp_j in TP **do**
    if tp_j covers rs_i **then**
      $RMS[rs.id].add(tp_j)$
    end if
  end for
end for

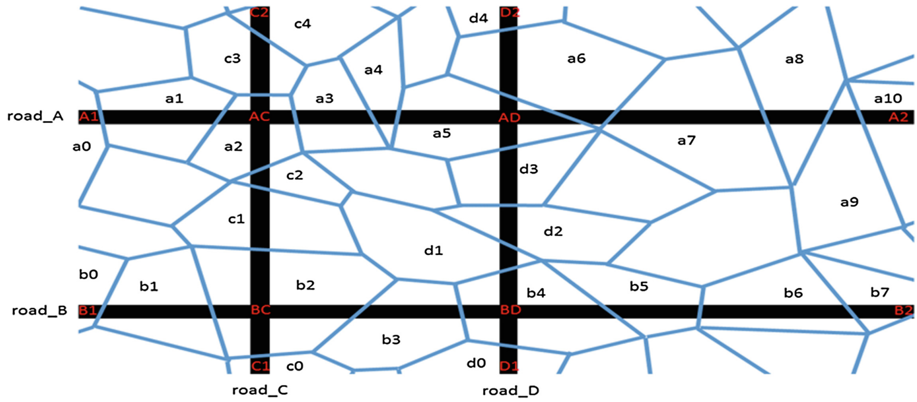

Fig. 2. Urban road network and its coverage base stations.

Know about the generation of the RMS, we demonstrate some road mobile sequences (RMS) in Fig. 2. Obviously, all the urban roads can be representated by mobile sequence. Figure 3 is an example which replaces roads with mobile sequences.

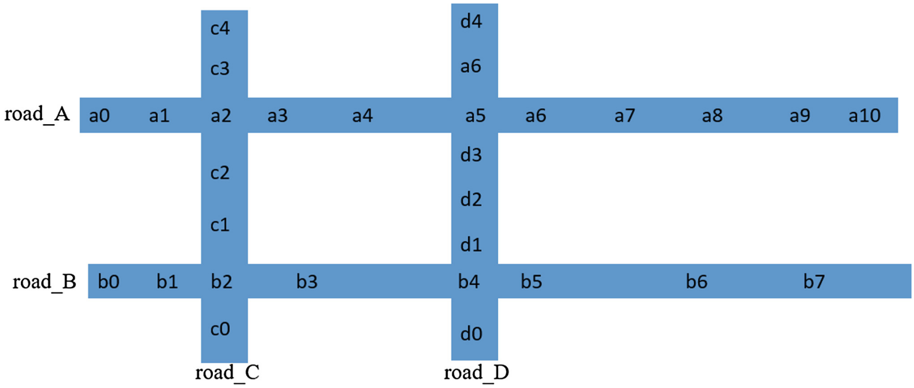

Fig. 3. RMS represents road network.

4.2 Extracting Human Trace Representation Based on Mobile Sequence (PMS)

Assume that base stations' identifiers are as a set BS_SET and the one which communicates with the users is identified by p ($p \in BS\_SET$). In some specified time span (t_s, t_e) $(t_s < t_1 < t_2 < \cdots < t_i < \cdots < t_n < t_e)$, where, $i = 1, 2, 3, \cdots, n$. A person's trace can be represented as (t_s, t_e). Now we take Fig. 2 as an example to describe how to generate PMS from mobile network logs. An assumption demonstrates like this, if a mobile terminal moves along $[A1, AC, AD, BD, D1]$ and time is ignored temporarily, then a possible log record is $[a0, a1, a1, a1, a2, a2, a3, a3, a3, a4, a5, a5, a5, d3, d3, d2, d1, b4, b4, b4, d0]$. Based on the log record, a mobile sequence can be obtained by merging the adjoining labels. Figure 4 illustrates the generation of mobile sequence.

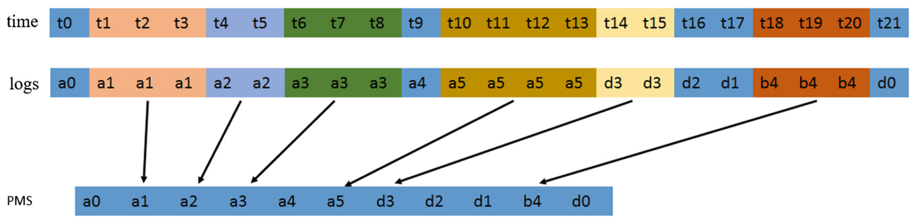

Fig. 4. The generation of PMS.

4.3 Obtain Human Trace by Matching PMS with RMS

As for the entire urban road networks, we can generate all the RMS and replace the representation form of the urban road networks with RMS network. Most of the roads can be described by RMS and most part of the people's urban traces can also be depicted by PMS. As a result, people's urban traces are converted to mobile sequence network traces. How strong the capability of the representation method will be expressed in Sect. 5. In this subsection, we solve the last problem in our method, that is how to obtain human traces by matching PMS with RMS in urban real roads network. Here is the algorithm (Algorithm 2). Find all parts of the RMS via traversing each point in PMS, then filter the points that has been added because of the intersection of two roads.

Algorithm 2. Obtain human trace by matching PMS with RMS Algorithm

Require: RMS=$\{\{ID_1,[rms_1,rms_2,...,rms_{n_1}]\},\{ID_2,[rms_1,rms_2,...,rms_{n_2}]\},\ldots,\{ID_k,[rms_1,rms_2,...,rms_{n_k}]\}\}$
Require: PMS=$\{pms_1,pms_2,\ldots,pms_m\}$
$HT \Leftarrow \{\}$; {//save human traces}
$threshold \Leftarrow 2$;
for pms_i in PMS **do**
  for rms_j in RMS **do**
    if rms_j contains pms_i **then**
      $HT[rms_j.id].add(rms_j)$;
    end if
  end for
end for
for ht_j in HT **do**
  if $length(ht_j) < threshold$ **then**
    $HT[rms_j.id].remove()$;
  end if
end for

5 Experiment

In previous section, we propose the RMS to represent urban real roads network and extract PMS from mobile network logs. However, the capability of the method should be measured and the data source we extract human traces also should be described. So the next subsection data source will be shown and some measurement will be defined to make sure the approach's representative capability. Therefore, in our matching procedure, an experiment will be undertaken. In our experiment, we define a measurement called CMR (correctly matching rate) and CR (covering rate) to measure the representative ability of the mobile sequences. Also, we use application to collect urban real roads information to validate if the RMS representation is effective in urban human trace mining.

5.1 Data Source

In our data analysis and experiment, three datasets will be used and some of them are from China Mobile Company. The first dataset is called mobile network logs, which is originated from users' network log records. The time span of the data is 7 days from 2016-05-03 to 2016-05-09 and the total size is 4.9 TB. Mobile network data are a kind of XDR data. A XDR record is generated by a session and uniquely indicated by a XDR identifier. When A session is transferring in the same base station, it generates a XDR record. When the session switches to a new base station, it generates a new XDR record. As for the long sessions, a time threshold is set to ensure a certain recording frequency. When a session time exceeds the threshold, the current record reaches its end and a new record starts. In XDR data, the threshold is 5 min, which is enough for traces.

Table 1. Partial fields of mobile network logs

Fields	Description
XDR_ID	Record id, the unique indicator of the record (coded)
IMSI	User id, the unique indicator of one user (coded)
IMEI	Phone id, the unique indicator of one phone (coded)
CELL_ID	A base station's id
PROCEDURE STARTTIME	A request start time, format: yyyy-MM-dd HH:mm:ss.S
PROCEDURE ENDTIME	A request start time format: yyyy-MM-dd HH:mm:ss.S
CELL_ID	A base station unique identifier
LONGITUDE	The base station longitude
LATITUDE	The base station latitude

The second dataset is called base station information, which contains 24905 records. All the location information describes Heping district, Tianjin City. The third dataset is urban real road network information. Every road is illustrated by a sequence and every point in a sequence have three variables: longitude, latitude and the road width at that point. The detail of the data is described in Table 1.

5.2 Measurement Definition

Now we define two measurements called CMR (correct matching rate) and CR (covering rate) to measure the representative ability of the theoretical RMS.

Definition 6: ***CMR (correctly matching rate).*** *CMR is used to describe the matching degree between the theoretical RMS and the practical RMS. We define the former as* $RMS_t = \{BS_{t1}, BS_{t2}, \cdots, BS_{ti}, \cdots, BS_{tn}\}$, *where* BS_{ti} *is the i-th base station in the theoretical sequence. The latter is defined as* $RMS_p = \{BS_{p1}, BS_{p2}, \cdots, BS_{pj}, \cdots, BS_{pm}\}$, *where* BS_{pj} *is the j-th base station in the practical sequence. If we find* n_1 *elements in the sorted set* RMS_p *and the found elements have the same order with* RMS_t, *then we define* $CMR = n_1/n$. *The larger CMR is, the stronger matching ability of the theoretical RMS and the practical RMS.*

Definition 7: ***CR (covering rate).*** *CR is used to demonstrate the covering ability of the theoretical mobile sequence. If a practical mobile sequence is covered by a theoretical mobile sequence perfectly, then we maintain that theoretical mobile sequence has strong representation ability to depict human trace. We define two sorted sets RMS_t and RMS_p as Definition 4. If we find m_1 elements in the sorted set RMS_p and the found elements have the same order with RMS_t, then we define $CR = m_1/m$. The larger the CR is, the stronger covering ability of the theoretical RMS.*

From the definitions above, CMR and CR are used to describe the match degree of RMS_p and RMS_t. If n, n_1 in definition CMR is equal with n, m_1 in definition CR, then it indicates that RMS_p and RMS_t match completely. As a matter of fact, our final purpose is to find out people's trace according to PMS, therefore, we define matching accuracy (MA) to describe the matching degree of PMS and RMS_t.

Definition 8: ***MA (matching accuracy).*** *MA is used to demonstrate the matching ability of the PMS and RMS_t. If a PMS is covered by a RMS_t perfectly, then we indicate that RMS_t has strong representation ability to depict PMS. We define two sorted sets RMS_t as Definition 4 and PMS (defined as $PMS = \{BS_{p1}, BS_{p2}, \cdots, BS_{pk}, \cdots, BS_{pl}\}$, where BS_{pk} is the k-th base station in the practical sequence). If we find l_1 elements in the sorted set PMS and the found elements have the same order with RMS_t, then we define $MA = l_1/l$. The larger the MA is, the stronger matching ability of the theoretical RMS.*

5.3 Real Roads Matching Theoretically Analysis

The relationship between the base station coverage area and roads (Fig. 1, some urban roads are marked as yellow) is analyzed as follows.

- Most of the base stations are usually distributed on both sides of the road. The road is normally covered by more than one base stations.
- In urban areas, base stations' density is relatively higher, the coverage of each base station is smaller (coverage radius is around 100 m).
- Also, to ensure running vehicle communication quality, the base station will be built along the main road.

Based on the above analysis, the road matching method can be obtained as follows:

- A user is moving a longer distance and produces a longer mobile sequence that can be matched with the road network. And the elements in the sequence (represented by the base stations) are located in both sides of the road.
- If the mobile sequence generated by the movement of a known user can matches the road network and the incremental mobile sequence still matches the road network when the user continues moving, then we judge that the user is on the road. If the generation of mobile sequence stops for a long time, then we stop matching.

– Because of some factors (weather, alternating day and night, etc.), a very small part of base stations' coverage will slightly get changed. So some threshold should be added in matching. If the distance is less than a threshold value we provide, then it can be thought correctly matched.

5.4 Using Application to Collect Real Roads Data

To determine whether the real world practical RMS matches the theoretical RMS generated by Thiessen polygon dividing the real roads, we developed an application program to collect the real world urban roads data and base stations information along the roads. We install our application on our mobile phone and keep running on some main urban roads. We collect the roads longitude and latitude sequences and the base stations' identifiers. The data fields we collected are shown in Table 2.

Table 2. Application collection data fields

Fields	Description
TIMESTAMP	Time, format: yyyy-MM-dd HH:mm:ss.S
GPS_LONGITUDE	Road updating longitude
GPS_LATITUDE	Road updating latitude
CELL_ID	Base station unique identifier
LAC	Location area code

5.5 The Theoretical RMS and Practical RMS in Real Urban Roads

As illustrated in Fig. 1, the base stations are distributed along the two sides of the roads. We assign some uppercase letters to represent the roads for convenience. Hereinafter, roads will be described by letters. To demonstrate every road theoretical RMS (RMS_t) and practical RMS (RMS_p), we take one road as an example (Fig. 5).

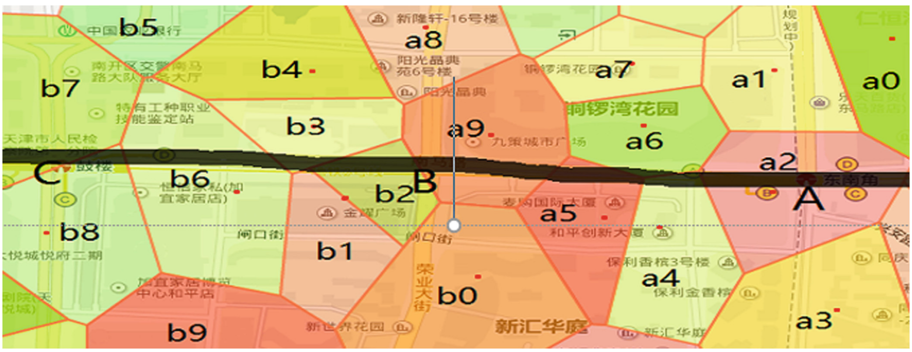

Fig. 5. An example to generate RMS_t and RMS_p in real roads.

First of all, to generate RMS_t, we visualized the dataset in Sect. 5.1 based on Algorithm 1, and part of the visualization is like Fig. 1. Then for all of the roads we assign some identifiers for them, respectively. Now we take Fig. 5 as an example to describe the case. In this figure, we use the combination of lowercase letters a–z and single digits 0–9 to identify a Thiessen polygon. Therefore, RMS_t is $\{a2, a4, a6, a5, a9, b2, b3, b6, b8, b7\}$ according to the definition of RMS. Meanwhile from our collected data, RMS_p is $\{a2, a6, a9, b3, b6, b8, b7\}$, which is generated based on the method in previous section.

5.6 Matching Results of RMS_t and RMS_p

After generating all of the experimental roads RMS_t and RMS_p, we demonstrate all the sequences in Table 3. From Table 3, we can see that RMS_p is shorter than RMS_t and almost all of the elements in RMS_p are contained in RMS_t, which means RMS_t has an excellent covering ability. In the perspective of the definition of CMR, we obtain an average 73.6%. But in practical trace mining

Table 3. Matching results of RMS_t and RMS_p

Road	RMS_t	RMS_p	CMR	CR
ABC	a2, a4, a6, a5, a9, b2, b3, b6, b8, b7	a2, a6, a9, b3, b6, b8, b7	0.700	1.000
BJ	a1, a3, a5, b5, b4, c4, c5	a1, a3, a5, b5, c4, c5	0.857	1.000
CDEF	a1, a3, a6, a7, a8, b2, b4, b5, b7, b9, c0, c1, c3, c4, c6, d0, c9, d5, d6	a1, a3, a7, a8, b2, b4, b7, b9, c0, c1, c4, c6, d0, d5, d6	0.789	1.000
DJK	a0, a2, a3, a4, a6, a8, b0, b4, b7, b9, c3, c4, c2, c5, c8	a0, a4, a6, a8, b0, b4, b7, c4, c2, c8	0.667	1.000
FOXWV	a0, a2, a3, a5, a7, a8, b0, b2, b3, b5, b6, b7, b9, c1, c0, c3, c4, c6, c8, d1, d2, d3, d7, d6, d9, e0, e2, e3, e4, e7	a0, a2, a3, a5, a7, b0, b2, b3, b7, c0, c3, c6, c8, d1, d4, d3, d7, e0, e2, e3, e7	0.700	0.952
KLM	a2, a6, a9, b3, b7, b9, b8, c2, c7	a2, a6, a9, b3, b9, b8, c2, c7	0.889	1.000
LNOPH	a0, a6, a5, a9, b3, b5, c1, c4, c7, c8, c9, d1, d2	a0, a5, b0, b3, b5, c1, c4, c7, d0, d1, d2	0.692	0.818
MYXQ	a0, a1, a6, a7, a8, c3, c4, d0, d1, d7, d6, e0, e1, e3, e7, f1, f4, g3, g2, g1, g7, h1	a0, a1, a7, a8, c3, c4, d1, d6, e1, e3, e7, f5, f4, g2, g1, h1	0.682	0.933
QR	a8, b2, b5, b4, b3, b9	a8, a7, b5, b4, b9	0.667	0.800
SW	a5, a6, a8, b1, b4, b6, b8, c2, c6, c7, c9, d1	a5, a6, a8, b4, b6, c2, c6, c7, c9, d1	0.833	1.000
TUV	a3, a6, a7, a9, b1, b4, b8, c4, c5, c6, c7, c8, d4, d7, d9	a3, a6, a7, b1, b4, b8, c5, c6, c8, d4, d5, d9	0.733	0.917
GPQST	a2, a3, a4, a6, a7, a8, b8, b7, c2, c3, c4, c5, c6, e5, e9, f6, f2, f3, f7, f8, g1, g2, g5	a2, a3, a4, a7, a8, b7, c2, c3, c5, c6, e5, e9, f6, f8, g2, g5	0.696	1.000
GHI	a2, a1, a3, a4, b1, b6, b5, b4, c7, c8, d0, c9, d5, d4, d7	a2, a3, a4, b1, b5, b4, b3, c8, c9, d5, d7	0.667	0.909
Mean			0.736	**0.948**

system or platform, we just need to consider whether the RMS_p is covered by RMS_t and whether the order of the sequences are consistent. So we provide a new measurement CR and obviously we get a decently higher CR. That is to say, mobile sequence is an effective approach to mine persons' mobile traces.

5.7 Matching Results of PMS and RMS_t

Now we use PMS to match RMS_t. PMS is extracted from mobile network logs utilizing the method described in Sect. 4.2 and RMS_t is the same as the Sect. 5.5. We use the same labeling method in Sect. 5.5 and give the points of PMS and RMS_t. Then the results is described in Table 4. From the results, we can see, mean MA is high, which means PMS matches RMS_t well. That is to say we can find a person's real roads according to PMS. Because PMS and RMS_t have mapping relationship and RMS_t and real roads network also have mapping relationship. From the mappings, human traces can be obtained.

Table 4. Matching results of RMS_t and PMS

Num	RMS_t	PMS	MA
1	a0, a1, a3, a4, a6, a8, b0, b2, b4, b7, b9, c3, c4, c5, c8	a1, a4, a8, b0, b4, b7, c2, c4, c8	0.889
2	a1, a2, a3, a5, a8, b0, b2, b3, b5, b8, b9, c0, c2, c3, c5, c7, d0, d1, d2	a1, a3, a5, b0, b3, b8, c0, c3, c7, d1	1.000
3	a0, a3, a5, b2, b3, b4, b6, c0, c2	a0, a3, b2, b4, b5, b6, c2	0.857
4	a0, a2, a1, a4, b1, b2, b4, b5, b8, b9, c0, c1, c3, c4, c7, c8, c9, d1, d2	a0, a1, b0, b2, b5, b9, c1, c4, c8, d1, d2	1.000
5	a0, a1, a2, a3, a4, a6, a8, a9, c1, c2, c3, c4, c6, c8, c9	a0, a1, a4, a8, c0, c2, c4, c6, c9	0.889
6	a2, a3, a9, b0, b3, b5, b8, b9, c1, c3	a2, a3, a4, a9, b0, b3, b8, c0, c3	0.778
7	a3, a4, a7, b0, b1, b3, b5, c0, c3, c5, c6, c8, d1, d3, d4, d6	a3, a4, a6, b1, b5, c5, c6, d1, d3, d5, d6	0.818
8	a2, a3, a5, a8, b0, b2, b9, c0, c3, c4, c7, c9, d2, d3, d4, d6, d7, d9	a2, a5, b0, b2, b9, c3, c9, d2, d4, d6, d9	1.000
9	a0, a2, a4, a5, a7, a9, b2, b7, c1, c2, c3, c4, c5, d0, d3, d4, d6, d7, d8, e2, e4, e5, e7, f0, f3, f4, f5, f7, g1, g2, g4, g5, g7	a0, a4, a7, a8, b2, c2, c5, c6, d3, d5, e2, e5, e9, f4, f7, g2, g5	0.824
10	a2, a3, a4, b0, b1, b3, b4, b6, b8, d0	a2, a3, b1, b4, b8, d0	1.000
Mean			**0.906**

6 Conclusion and Future Work

In this paper, we proposed a novel representative method called mobile sequence to depict urban roads and then human's mobility traces can be described and mined more easily and conveniently. Then we collect real urban roads information and urban base stations identifier to validate the effectiveness of our method, as a result, they gain a very high CMR and MA, which means the method maintain a strong capability of representation urban traces. That is to say, on one hand we just need to extract PMS from mobile network logs and establish the mapping relationship with RMS_t. On the other hand, we also build the mapping relationship between RMS_t and urban real roads network. From the analysis and experimental results, the method we proposed to represent urban traces is excellent in mining urban trace.

In mobile sequence, we have a bottleneck. That is if two mobile traces or urban roads share a common mobile sequence, it is hard to determine which road the person is on. On one hand, this case happens because of some tiny roads locate along the main urban roads. The efficient areas of mobile sequence are generated by Thiessen polygon, so to solve tiny roads problem is a little difficult. But on the other hand this problem can be partially solved based on trace frequency and real urban roads network in practical urban trace query platform in that we have human's network data as well as the urban roads network data. From the distinct data source, we will probe a more fine-grained urban trace mining approach in the future work. Meanwhile, the complexity of the proposed algorithm is high and there is much space to improve the matching algorithm to reduce the computing complexity though we can handle this problem on spark platform.

Acknowledgments. This work is supported by Ministry of Education-China Mobile Research Fund under grant MCM20150507 and Tsinghua University Initiative Scientific Research Program (No. 20131089190). Beijing Key Lab of Networked Multimedia also supports our research work.

References

1. Aurenhammer, F.: Voronoi diagrams—a survey of a fundamental geometric data structure. ACM Comput. Surv. **23**, 345–405 (1991)
2. Brockmann, D., Hufnagel, L., Geisel, T.: The scaling laws of human travel. Nature **439**(7075), 462–465 (2006)
3. Caragliu, A., Del Bo, C., Nijkamp, P.: Smart cities in Europe. J. Urban Technol. **18**, 65–82 (2011)
4. Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S., Trasarti, R.: Unveiling the complexity of human mobility by querying and mining massive trajectory data. VLDB J. Int. J. Very Larg. Data Bases **20**(5), 695–719 (2011)
5. Giannotti, F., Nanni, M., Pinelli, F., Pedreschi, D.: Trajectory pattern mining. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 330–339. ACM (2007)
6. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. Nature **453**(7196), 779–782 (2008)
7. Hasan, S., Schneider, C.M., Ukkusuri, S.V., González, M.C.: Spatiotemporal patterns of urban human mobility. J. Stat. Phys. **151**(1–2), 304–318 (2013)
8. Iqbal, M.S., Choudhury, C.F., Wang, P., González, M.C.: Development of origin–destination matrices using mobile phone call data. Transp. Res. Part C: Emerg. Technol. **40**, 63–74 (2014)

9. Lane, N.D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., Campbell, A.T.: A survey of mobile phone sensing. IEEE Commun. Mag. **48**(9), 140–150 (2010)
10. Luo, W., Tan, H., Chen, L., Ni, L.M.: Finding time period-based most frequent path in big trajectory data. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, pp. 713–724. ACM (2013)
11. Okabe, A., Boots, B., Sugihara, K.: Spatial Tessellations: Concepts and Applications of Voronoi diagrams. Wiley, Hoboken (1992)
12. Simini, F., González, M.C., Maritan, A., Barabási, A.L.: A universal model for mobility and migration patterns. Nature **484**(7392), 96–100 (2012)
13. Song, C., Koren, T., Wang, P., Barabási, A.L.: Modelling the scaling properties of human mobility. Nat. Phys. **6**(10), 818–823 (2010)
14. Song, C., Qu, Z., Blumm, N., Barabási, A.L.: Limits of predictability in human mobility. Science **327**(5968), 1018–1021 (2010)
15. White, J., Wells, I.: Extracting origin destination information from mobile phone data. In: 2002 Eleventh International Conference on Road Transport Information and Control (Conf. Publ. No. 486), pp. 30–34. IET (2002)
16. Wu, C., Xu, B., Li, Q.: Parallel accurate localization from cellular network. In: Wang, Y., Xiong, H., Argamon, S., Li, X.Y., Li, J.Z. (eds.) BigCom 2015. LNCS, vol. 9196, pp. 152–166. Springer, Cham (2015). doi:10.1007/978-3-319-22047-5_13
17. Wu, C., Xu, B., Shi, S., Zhao, B.: Time-activity pattern observatory from mobile web logs. Int. J. Embed. Syst. **7**(1), 71–78 (2014)
18. Zhang, C., Han, J., Shou, L., Lu, J., La Porta, T.: Splitter: mining fine-grained sequential patterns in semantic trajectories. Proc. VLDB Endow. **7**(9), 769–780 (2014)
19. Zheng, Y.: Trajectory data mining: an overview. ACM Trans. Intell. Syst. Technol. (TIST) **6**(3), 29 (2015)
20. Zheng, Y., Capra, L., Wolfson, O., Yang, H.: Urban computing: concepts, methodologies, and applications. ACM Trans. Intell. Syst. Technol. **5**, 38 (2014). http://research.microsoft.com/apps/pubs/default.aspx?id=211950
21. Zheng, Y., Zhou, X.: Computing with Spatial Trajectories. Springer Science & Business Media, New York (2011)