Prediction of User Mobility Pattern on a Network Traffic Analysis Platform

Haiyang He, Yuanyuan Qiao, Sheng Gao, Jie Yang, Jun Guo
Research Center of Network Monitoring and Analysis
Beijing University of Posts and Telecommunications, China
{hehaiyang, yyqiao}@bupt.edu.cn

ABSTRACT

The mobile Internet brings tremendous opportunities for researchers to analyze user mobility pattern, which is of great importance for Internet Service Providers (ISP) to provide better location-based services. This paper focuses on predicting user mobility patterns based on their different mobility characteristics. For that, we collect real-world data from Long Term Evolution (LTE) mobile network by a specially developed network traffic analysis platform followed by clustering the user into stationary one or mobile one with a location-entropy-based method for distinguishing groups with distinct mobility characteristics, and then we present the tailored Intelligent Time Division (ITD) method and Time-Based Markov (TBM) predictor for the location prediction of stationary and mobile users respectively. Extensive experiments demonstrate the effectiveness and better performance of our proposed methods compared with the baselines, as well as the adaptabilities of different predictors according to individual's mobility characteristics.

Categories and Subject Descriptors

C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design—Wireless communication

Keywords

LTE Network; Network Traffic Analysis; Mobility Pattern; Prediction Accuracy; User Clustering

1. INTRODUCTION

Recent years have witnessed the rapid development of mobile Internet and an exponential growth of mobile traffic data, which makes it possible to analyze user behavior in a large scale. With mobile traffic data, we can investigate many important features like data usage, mobility pattern and application usage [12]. Additionally, mobile traffic data is more fine-grained both temporally and spatially compared to Wireless Local Area Network (WLAN) and Call Detail Records (CDR) data [13]. In mobile network, a user's information is updated even if he/she is in the period of no data network activity, which ensures that the collected user trajectories are more continuous than CDR data. In addition, urban range data can overcome the limit of small dataset based on WLAN technology.

In order to monitor and analyze traffic data, a number of tools have been developed. For example, *tcpdump*[1] , *wireshark*[2] and *netflow*[3] provide interactive user interface for analysts to analyze captured packets, which made it broadly used in commercial networks. Over the years, as networks grow exponentially, traditional network systems can no longer handle the huge traffic data. Hadoop is an effective and popular platform to conduct large scale data. Many big companies like Yahoo, Facebook and IBM are using Hadoop for web log analysis and machine learning. In [7], the authors developed a Hadoop-based tool and made the first attempt to process Netflow data in Internet. However, there still lacks platform to systematically monitor and analyze mobile traffic data. In this paper, we propose a platform to conduct mobile traffic data.

Among various kinds of mobile traffic data analysis, understanding human mobility pattern and predicting their locations are of great importance in many areas. By extracting user mobility pattern, network operators can provide efficient network planning for mobile phone users and then develop more reliable communication protocols [8]. If an application is furthermore equipped with a function of location prediction, it can provide personalized services.

Previous researchers have studied several mobility prediction models like Markov models [2], LZ-based methods [3,6] and Bayesian approaches [1]. Among these models, Markov and LZ-based methods get more attentions because of their low computational complexity and resource needs. In [11], Researchers assessed several algorithms and found the simple low-order Markov predictors worked as well or better than the other predictors.

Aforementioned prediction methods consider spatial context as main factor to affect user mobility prediction. Recently, more and more researchers begin to consider temporal factors. The authors in [5] demonstrated that time significantly impact randomness, size, and probability distribution of people's movements. The authors of [10] proposed an algorithm called NextPlace which used arrival time and

MobiArch'15, September 7, 2015, Paris, France.

DOI: http://dx.doi.org/10.1145/2795381.2795389.

[1]http://www.tcpdump.org
[2]https://www.wireshark.org
[3]http://www.cisco.com/c/en/us/products/ios-nx-os-software/ios-netflow/index.html

duration information to predict future locations. In [4], the author clarified two prediction question which were "where will a person be at a certain time" and "when will a person be at a certain place". These methods demonstrate that time-based mobility predictions are more effective compared to the case when temporal factor is not considered.

Besides spatial and temporal factors, in this paper, we also consider user's mobility characteristic has a significant impact on how we predict user mobility pattern. For example, if we want to predict where a user will be within a period of time, it is easier to predict a white collar that has a nine-to-five job than to predict a salesman who runs the business every day, since a white collar has a more regular life. Our previous study [9] suggested that we should consider applying different prediction methods on the users with distinct spatial-temporal characteristics. In this paper, we provide a further research on predicting user mobility patterns based on their different mobility characteristics.

The key contributions and some interesting findings of this paper are as follows:

- We propose a network traffic analysis platform to conduct mobile traffic data. The platform is designed and implemented following a multilevel architecture with collecting module, distributing module, batch & stream processing module, interface module, and a cluster manager.

- For mobility prediction, we consider users' mobility characteristics, spatial and temporal factors as main features that may affect user mobility patterns. We use an entropy-based method to cluster users into stationary or mobile users. For different types of users, we then apply two different time-based methods to predict their mobility patterns.

- We find that the Intelligent Time Division (ITD) method is more effective for stationary users. The average prediction accuracy of ITD can get 71.0% compared with 56.9% without time considered. For mobile users, we propose the Time-Based Markov (TBM) predictor to predict their next hop and arrival time. The prediction accuracy of TBM can be finally improved to 59.2% while the baseline accuracy of Markov is 46.8%.

The remaining of the paper is organized as follows. The next section introduces the architecture of our traffic analysis platform. Methods for user clustering and mobility prediction are introduced in section 3. Then, section 4 will show our experimental results. Finally, conclusion and future work are presented in Section 5.

2. TRAFFIC ANALYSIS PLATFORM

Figure 1 shows the architecture of our self-developed network traffic analysis platform. The platform could be deployed in 2G, 3G and 4G networks. An exposition of its major components is offered next.

Cellular Network

There are three major components in cellular network: User Equipment (UE), Radio Access Network (RAN) and Core Network (CN). UE is the device that an end-user uses to connect with cellular network. RAN establishes the connection between UE and CN. Each RAN consists of several transceiver stations (called BTS, Node-B, and eNodeB respectively in 2G, 3G and 4G networks). In 2G/3G core network, the Serving GPRS Support Node (SGSN) establishes a tunnel with a Gateway GPRS Support Node (GGSN) that provides connectivity to Internet. In LTE core network, the components that are used to connect Internet are composed of Serving Gateway (SGW) and PDN Gateway (PGW).

Collecting Module

Our platform monitors the network traffic by Traffic Monitoring System (TMS) developed by our research team. The TMS is deployed between RAN and CN. It has been deployed in the production networks by several ISPs for traffic monitoring purposes. TMS capture all downlink and uplink IP packets, and generate the flow records. All flow records are collected by Collector, and transferred to distributing module.

Distributing Module

The distributing module reads flow records from collector, and distributes the data to batch & stream processing module. The distributing module provides load balancing, fault tolerance and good scalability.

Batch & Stream processing Module

Batch & Stream processing module is a Hadoop-based System. Batch processing module is responsible for offline analysis. In order to store a huge size of traffic data, we build a distributed data store based on Hadoop Distributed File System (HDFS) and Hadoop Database (HBase). Analyzer is a sequence of MapReduce jobs that can run automatically every midnight. Additionally, Hive and Pig provide Hive-QL and Pig Latin for data analysts to analyze the data more conveniently. Stream processing module is composed of Kafka and Storm, among which Kafka is a high-throughput messaging system for real-time data, and Storm is a distributed real-time computation system. The usage of these two open source tools makes our real-time computing application easier to develop.

Interface Module

The analysis results produced by Batch & Stream processing module are saved to relational database PostgreSQL. In order to present the analysis results in a user-friendly manner, we provide web Graphical User Interface (GUI). At the same time, analysts can write and execute Pig scripts through the web directly.

Cluster Manager

The cluster manager is a self-developed software which can monitor software and hardware on each server. It can collect performance metrics of servers such as Central Processing Unit (CPU), memory, disk and network utilization. Whenever an error or warn occurs, the cluster manager will send alerts to admin via emails and Short Messaging Service (SMS), then the admin can view the detailed error information and manage the cluster through the web GUI.

Based on the network traffic analysis platform, we can collect mobile traffic data and analyze user mobility. In this paper, we use a dataset collected from a Chinese LTE service provider from October 10, 2013 to October 31, 2013. In LTE network, the closest eNodeB is recorded when a user connects the Internet with his/her mobile device. Even if the user has no network activity, the information of the user will still be updated every 12 minutes. The collected dataset is composed of a sequence of fields such as userID, timestamp, latitude & longitude (eNodeB's location) and signaling procedure Code, and there are 3474 users and 729

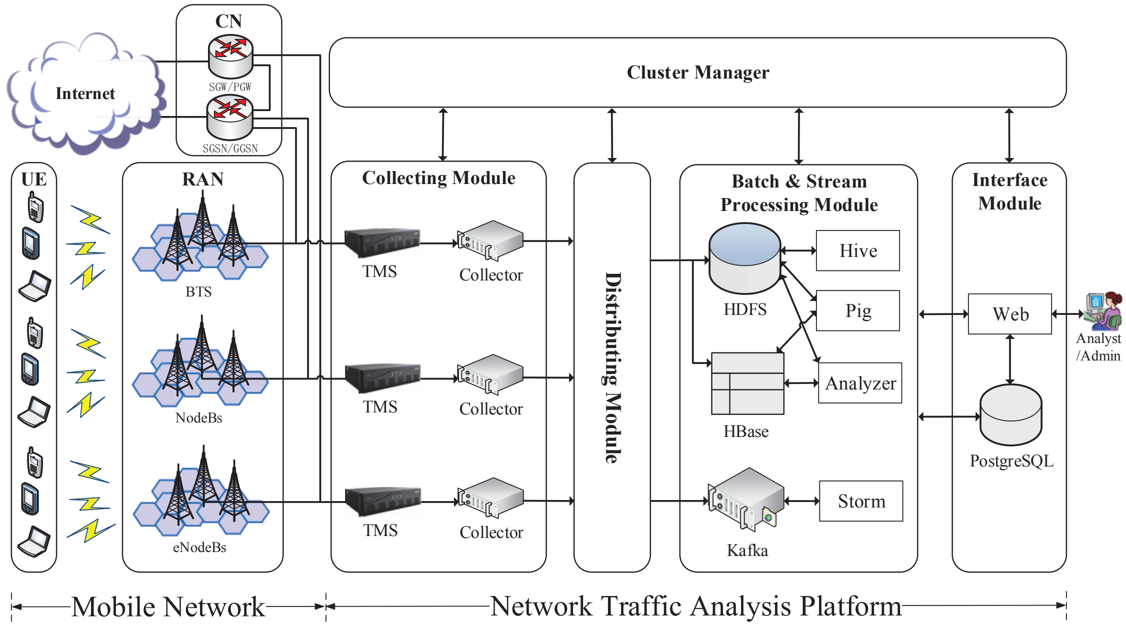


Figure 1: Architecture of network traffic analysis platform.

eNodeBs in the three-week dataset. Considering some users never move or have few records (often cut off the network) in the dataset, which makes it impossible to analyze their mobility pattern, we select 2204 users who connect to more than 10 eNodeBs and connect more than 500 times. For the security reason, users' privacy information is replaced by a hashed number, which could be used for marking users without affecting the results of our prediction.

3. METHODS FOR USER CLUSTERING AND MOBILITY PREDICTION

In this section, we first introduce a location-entropy-based method for user clustering, and then apply two methods on stationary and mobile users respectively.

3.1 Location-entropy-based User Clustering

In order to cluster users according to their mobility characteristics, we introduce a location-entropy-based method. If a user connects to M base stations for T times in all, and the times he connects to the m_{th} base station is t_m, $0 < m < M$, then the location entropy can be calculated as follows:

$$H(L) = -\sum_{m=0}^{M} \frac{t_m}{T} log \frac{t_m}{T} \tag{1}$$

Location entropy captures the mobility activity of a user, and shows us how predictable a user's location is. The higher the location entropy value is, the more unpredictable of the user is.

Based on the definition of location entropy, the clustering method can be described as follows:

Step 1: Discretize day into 24 time intervals, each of which lasts one hour long.

Step 2: For each user, calculate the entropy in each time interval, and get an entropy vector that contains 24 entropy values.

Step 3: Use the entropy vectors of each user to cluster users with k-means clustering.

Note that if a user often moves, the entropy will keep in a high value in his/her entropy vector. By applying this method, we will cluster users into two groups, i.e. stationary and mobile users. Here, the main difference between stationary and mobile users is the number of base stations they connect but not related to the base stations' positions. If a user connects to very few base stations in each time period, no matter how far distance he/she travels, he/she will be treated as a stationary user, and we will try to predict his/she locations with a given time. On the other hand, if a user often switches among different base stations, even though he/she moves around within a small community, he/she will be treated as a mobile user, and we will try to find his/her moving trajectory.

3.2 ITD Method for Stationary Users

Stationary users usually move in a particular period of time during a day, such as commute time of white collars, shopping time for a housewife. However, we can't define a pattern that every stationary user follows. For example, students have to sleep early and get up early to go to school while most of computer programmers prefer working at night and get to work late in the morning, so the mobility pattern of a student differs from that of a programmer.

The "intelligent time divisions" (ITD) method [5] takes spatial probability distribution as a significant characteristic to predict user mobility pattern. If taking time factor into consideration, we can define the spatial probability distribution as

$$P_t(X,Y) = prob(\, x(t) = X \;\&\; y(t) = Y) \tag{2}$$

where $(x(t), y(t))$ represents the location of a user at time t. Spatial probability distribution of a user shows the probability that a user is at a particular location which is useful for mobility prediction.

Here, we tailor the ITD method as below:
Step 1: For each user, divide time into 24 time intervals, each of which lasts one hour long.
Step 2: Calculate spatial probability distribution for each time interval to obtain hourly distributions. The data form of each distribution is a matrix, with row & column representing longitude & latitude.
Step 3: Cluster the above 24 intervals based on agglomerative hierarchical clustering. Each cluster includes time intervals that have similar spatial probability distributions.
Step 4: Use spatial probability distribution of the clusters to predict user's future location. In each cluster, the main location that has the maximum spatial probability will be the predicted location in these time intervals. If a user appears n times in all and m times are predicted correctly, we have $accuracy = m/n$.

Based on ITD method, we can find out the mobility patterns of stationary users, and then predict their future location within a given time period.

3.3 TBM Algorithm for Mobile Users

Unlike stationary users, mobile users always move from one place to another, like postman, driver or traveler, so it is difficult to predict their location correctly with a given time. However, we may find patterns from their history trajectory to predict their next hop. Algorithms on predicting next hop of a user like Markov and LZ-based predictors are independent of time. Here we propose a Time-Based Markov (TBM) algorithm.

We define (t_i, loc_i) if a user visit loc_i at time t_i. Based on user's history trace $((t_1, loc_1), ..., (t_n, loc_n))$, if the user's current time and location $(tr, locr)$ are known, TBM will predict his/her next hop (i.e., $nextr$), and the time interval between $locr$ and $nextr$ (i.e., dr). The main steps of TBM are as below.
Step 1: Create user's time and location series: $(T, LOC) = ((t_1, loc_1), \dots, (t_n, loc_n))$.
Step 2: Given a certain $(tr, locr)$ and a time threshold T_s, search for records (t, loc) from the series (T, LOC) that satisfies $loc = locr$ and $tr - T_s < t < tr + T_s$.
Step 3: If no record satisfies the above condition, NO MATCH is returned as results.
Step 4: If there are $m (m > 0)$ records meet the demand, we name them $((tr_1, locr_1), \dots, (tr_m, locr_m))$. Then we search for each $locr_j$'s next location $nextr_j$ from (T, LOC), and calculate the time intervals dr_j(round down to hour) between $locr_j$ and $nextr_j$ $(1 \leqslant j \leqslant m)$. Here we have a series $(nextr_1, dr_1), \dots (nextr_m, dr_m)$.
Step 5: Select $(nextr, dr)$ that owns the maximal probability.

Here we give an example to explain the algorithm. If it is 9:00 am, Tom is at location c, and we set $T_s = 1h$. Firstly, TBM searches for Tom's history trajectories as:

(7:10,a), (8:10,b), (8:40,c), (9:00,d), (10:00,e), (12:00,c)...
(7:40,a), (7:50,c), (8:30,d), (11:10,e), (12:40,b), (13:50,c)...
(7:30,a), (8:00,b), (9:00,c), (9:50,e), (12:00,b), (13:30,c)...
(8:20,b), (9:00,c), (9:30,d), (9:40,c), (10:00,e), (12:00,c)...
(7:30,a), (8:10,b), (9:10,c), (11:40,d), (12:00,c), (12:40,b)...
(7:10,a), (9:30,c), (10:20,d), (10:00,e), (10:10,d), (12:00,e)...

Then TBM selects records that match the condition: $loc = c$ and $8:00 < t < 10:00$. Six required records have been added shadows. Next, TBM finds these records' next records (added underline), and then extracts the candidates $(nextr, dr)$ with location and arrival time as: *((d,0), (e,0), (d,0), (e,0), (d,2), (d,0))*. Because $(d, 0)$ has the maximum probability 3/6, TBM predicts that Tom will move to location d in an hour.

4. EXPERIMENTAL RESULTS

In this section, we firstly implement the location-entropy-based method to cluster users into stationary or mobile users. The prediction results with ITD and TBM are shown in 4.2 and 4.3 respectively.

4.1 User Clustering Results

In order to choose an appropriate k for k-means method while clustering, we vary the values of k as $k = 2$, $k = 3$ and $k = 4$ respectively in our experiments, and find there is always a cluster has a very low entropy (e.g., low mobility). We define the users in this cluster as stationary user. Other clusters share the similar characteristic that has high entropy during the day-time. In order to simplify the model, we simply choose $k = 2$. The clustering result is shown in Figure 2. The number of stationary and mobile users are 795 and 1409 respectively. Based on the two types of users, we will conduct experiments on them respectively.

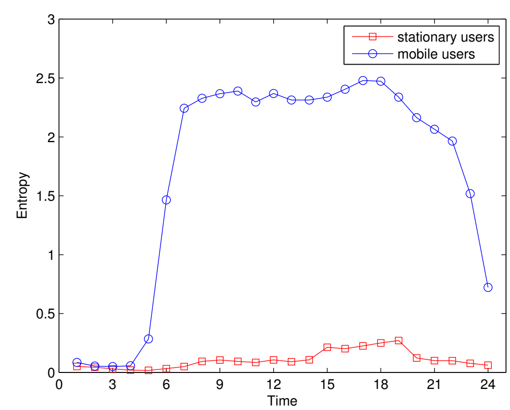
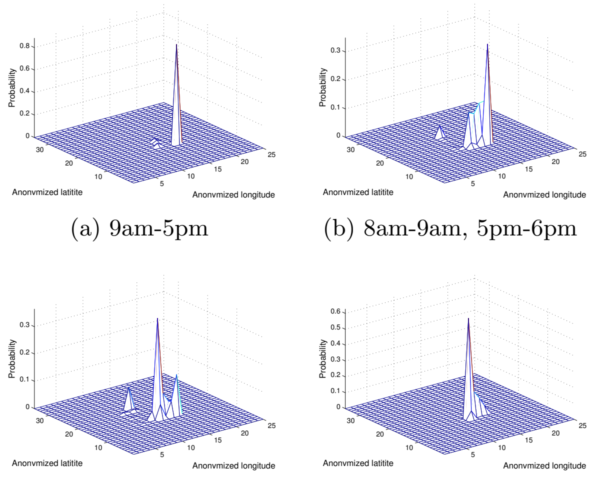


Figure 2: User clustering based on entropy.

4.2 Predicting Location of Stationary User

In this part, by applying the ITD method, we use the previous two-week data as training data, and predict users' locations in the third week.

Figure 3 shows a sample stationary user's spatial probability distribution of each time group. Latitude and longitude values are replaced to protect the user's privacy. As can be seen, the sample user tends to stay at a single place and seldom goes anywhere else from 11:00pm to 8:00am and from 9:00am to 5:00pm, which indicates that the two places are likely to be his/her home and working place. The places in Figure 3(b) are most likely between his/her home and working place, so we can infer that 8:00am-9:00am and 5:00pm-6:00pm are commute time periods for the user. From 6:00pm to 11:00pm, the user stays at limited places which is likely to be near his/her home. We use this pattern to predict his/her future locations. The prediction accuracy can achieve 86.2%,



(a) 9am-5pm (b) 8am-9am, 5pm-6pm

(c) 6pm-11pm (d) 11pm-8am

Figure 3: Pattern of a sample stationary user.

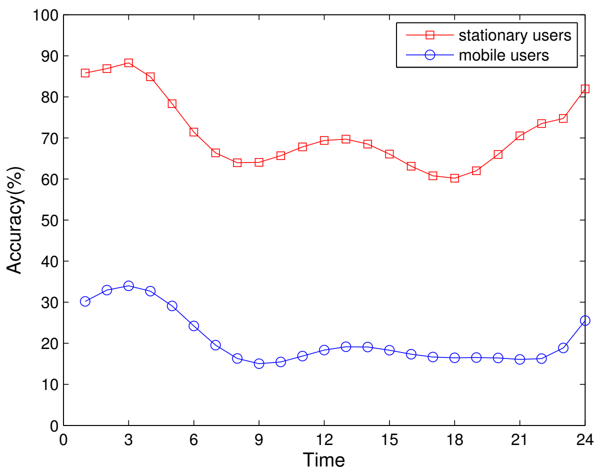


Figure 4: Prediction accuracy of ITD.

while the prediction based on a single spatial distribution that does not consider temporal factor gets 58.4% accuracy.

We apply the method to stationary and mobile users respectively, and the result is shown in Figure 4. Here we get time as horizontal axis and prediction accuracy as vertical axis. We can see the accuracy of stationary users is higher than mobile ones on the overall level, which suggest the method works better on stationary users. The peak of accuracy for both groups appears in the night when most people are sleeping at home. It is also worth noting that the accuracy has two low points for stationary users at 9:00am-10:00am and 5:00pm-6:00pm, which indicates the general commuting time falls into these two periods. The whole time average prediction accuracy of stationary users can achieve 71.0%, while prediction based on each stationary user's single spatial probability distribution without considering temporal factors gets only 56.9% accuracy. Taking time factors into consideration brings nearly 25% performance increment.

Table 1: Prediction accuracy of Markov & TBM.

Prediction accuracy	All users	Mobile users
Markov	38.2%	46.9%
TBM ($T_s = 1h$)	38.4%	47.7%
TBM ($T_s = 2h$)	39.2%	48.6%

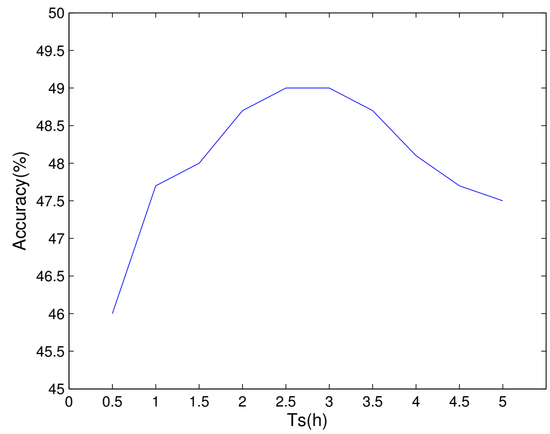


Figure 5: Influence of T_s on TBM.

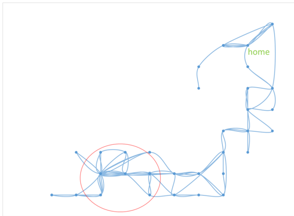
4.3 Predicting Next-hop of Mobile User

In this part, we use TBM algorithm to predict next-hop of mobile users. If a user connects to the same base station during a period of time, we merge all these records into a single record and the arrival time is set to the first record's time. We use the previous two-week data as training data, then if a user in the third week moves n times in all and m movements are predicted correctly, we have $accuracy = m/n$. Each user's accuracy will be averaged to get a final prediction accuracy.

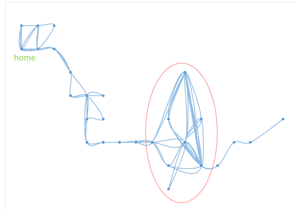
We first perform experiments with TBM and Markov on both mobile users and all users. For TBM predictor, we choose $T_s = 1h$ and $T_s = 2h$ respectively. The results are shown in Table I. As can be seen, both Markov and TBM perform better when predicting next-hop of mobile users. When $T_s = 1h$, the prediction accuracy of TBM predictor is improved from 46.9% to 47.7%, while $T_s = 2h$, the prediction accuracy goes higher.

In order to explore the influence of T_s on TBM, we set T_s vary from [0.5, 5] with step size of 0.5, and run the TBM algorithm with each T_s on mobile users. We get T_s as horizontal axis and accuracy as vertical axis shown in Figure 5. From the figure we can conclude that 2.5 to 3 hours is the best choice of T_s. This reveals that if we want to find a mobile user's pattern of a certain moment such as 10:00am, it is better to look at his/her movements from 7:30am to 12:30am every day.

We further examine TBM by focusing on the bad prediction results. We draw some trajectories of sample mobile users in Figure 6. In each subfigure, each point represents a base station, and the line between two points indicates that user has moved from one point to the other. The locations that are likely to be predicted wrong have been marked in a circle and we name them error-prone area. We find that users in error-prone area often have the certain pattern like



(a) Sample user 1



(b) Sample user 2

(c) Sample user 3

(d) Sample user 4

Figure 6: Trajectory of sample mobile users.

$loc_a - loc_b - loc_c - loc_d - loc_c - loc_b - loc_a - loc_b - loc_c - loc_d - loc_c - loc_b - loc_a$. Supposed that this is a trace of a user who is now at loc_b, the probability of moving to loc_a and to loc_c is nearly the same. This suggests that we should consider the previous location of loc_b. Thus, we change the 5_{th} step of our TMB algorithm: define $prevr$ to be the previous location of $locr$ and select $(nextr, dr)$ that has the maximal probability, then, if the selected $nextr = prevr$, select $(nextr, dr)$ that has the second maximal probability. After improving TBM algorithm this way, we can achieve 59.2% prediction accuracy for mobile users.

5. CONCLUSIONS

In this paper, we proposed a network traffic analysis platform, and collected real traffic data from LTE network in a city area of south China based on the platform. We then used a location-entropy-based method to cluster users with different mobility characteristics into stationary or mobile users. For stationary users, ITD method is more effective to predict their location within a period time. The average prediction accuracy of ITD can achieve 71.0%, while the prediction that does not consider temporal factor gets 56.9% accuracy. For mobile users, we proposed a TBM algorithm to predict their next hop and arrival time. We improved the prediction accuracy to 59.2% compared with 46.8% of basic Markov. Our experimental results suggest that ISP or Location Based Service Provider should consider applying different predicting methods to users with distinct mobility characteristics while predicting users' mobility patterns.

For the future work, some other special situations such as weekend and activity changes should also be considered to predict user mobility patterns. What's more, since the location-entropy-based method only considered the number of connections of a user to the base stations, which make the clustering result a little coarse, a better clustering method is expected. In addition, we used base station user currently connect as user's current location. However, if a user switches frequently among very close base stations due to cell oscillation, methods of removing data noise is expected.

6. ACKNOWLEDGMENT

This work was supported in part by National Natural Science Foundation of China (61273217, 61300080), Innovation Foundation of Beijing University of Posts and Telecommunications (2015RC11), 111 Project of China (B08004) and EU FP7 IRSES Mobile Cloud Project (612212).

7. REFERENCES

[1] S. Akoush and A. Sameh. Mobile user movement prediction using bayesian learning for neural networks. In *Proceedings of the 2007 international conference on Wireless communications and mobile computing*, pages 191–196. ACM, 2007.

[2] D. Ashbrook and T. Starner. Using gps to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 7(5):275–286, 2003.

[3] A. Bhattacharya and S. K. Das. Lezi-update: An information-theoretic framework for personal mobility tracking in pcs networks. *Wireless Networks*, 8(2/3):121–135, 2002.

[4] I. Burbey. *Predicting future locations and arrival times of individuals*. PhD thesis, Virginia Polytechnic Institute and State University, April 2011.

[5] S. Gatmir-Motahari, H. Zang, and P. Reuther. Time-clustering-based place prediction for wireless subscribers. *IEEE/ACM Trans. Netw.*, 21(5):1436–1446, Oct. 2013.

[6] K. Gopalratnam and D. J. Cook. Online sequential prediction via incremental parsing: The active lezi algorithm. *Intelligent Systems, IEEE*, 22(1):52–58, 2007.

[7] Y. Lee and Y. Lee. Toward scalable internet traffic measurement and analysis with hadoop. *SIGCOMM Comput. Commun. Rev.*, 43(1):5–13, Jan. 2012.

[8] A. J. Nicholson and B. D. Noble. Breadcrumbs: forecasting mobile connectivity. In *Proceedings of the 14th ACM international conference on Mobile computing and networking*, pages 46–57. ACM, 2008.

[9] Y. Qiao, J. Yang, H. He, Y. Cheng, and Z. Ma. User location prediction with energy efficiency model in the Long Term-Evolution network. *International Journal of Communication Systems*, 2015.

[10] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. T. Campbell. Nextplace: a spatio-temporal prediction framework for pervasive systems. In *Pervasive Computing*, pages 152–169. Springer, 2011.

[11] L. Song, D. Kotz, R. Jain, and X. He. Evaluating next-cell predictors with extensive wi-fi mobility data. *Mobile Computing, IEEE Transactions on*, 5(12):1633–1649, 2006.

[12] J. Yang, Y. Qiao, X. Zhang, H. He, F. Liu, and G. Cheng. Characterizing user behavior in mobile internet. *Emerging Topics in Computing, IEEE Transactions on*, 3(1):95–106, March 2015.

[13] Y. Zhang. User mobility from the view of cellular data networks. In *INFOCOM, 2014 Proceedings IEEE*, pages 1348–1356, April 2014.