Mining Users' Mobility Patterns Based On Apriori

Hongyan Cui[1,2,a], Xiaolong Yin[1,2,b]

[1]State Key Lab. of Networking and Switching Technology,
[2]Beijing Key Lab. of Network System Architecture and Convergence, Beijing University of Posts and Telecommunications, Beijing, China
[a]cuihy@bupt.edu.cn,[b]yxl198963@126.com

Keywords: Mobility patterns, Apriori Algorithm, Trajectory prediction, User behaviour analysis.

Abstract. Mobility pattern of users is decided by people themselves. However, in the daily life, mobile users often repeat regular routes in certain periods. To effectively mine the mobile rules of users and allocate the resource among mobile telecommunication operators and Internet Service Providers (ISP), grasping the models of human mobility can enable effective network planning and advertisement recommending. In this paper, we employ Apriori algorithm method to mine mobile users' movement model. The algorithm proposed is based on mining the mobility patterns of users, forming mobility rules from these patterns, and finally predicting a mobile user's next movement by using the mobility rules. The user mobility patterns are mined from the history of mobile user trajectories. In addition to the analysis of individual mobile behavior, it is possible to mine population movement regularity from a large number of user mobile patterns. By studying the group characteristics the place that a large number of users accessed can be found. Operators can optimize the setting of the base station according to the density of users to meet the needs of mobile users.

I. Introduction

Data collected from the base station of mobile phone users are used to analyze the social networks, mine user behavior [1,2]. This work mainly consists of the analysis of user's daily activities, building structural model of social relations, finding out the relationship between the individuals and the social networks, marking important position information. In addition, there is a corresponding application in the studies of the mobile user's location in AdHoc networks. [3,4] Wireless network-related researchers also became interested in the behavior of mobile user, and collected the data from the Wi-Fi environment for mobile user behavior modeling[5,6]. Gonzalez and others also use the data collected from base station for mobile user behavior modeling. Gonzalez's studies have shown that there are different laws in user's trajectories because of different times, different locations. For example, users tend to move between several important positions, spend three-fourths of time in this a few important positions. The most frequent path also appears between the several important positions [7]. Association rules are one of the main modes of the current data mining methods focusing on establishing links between the different fields of data. [8]

In this paper, an improved sequential pattern mining algorithm was adopted to find the frequent patterns of users which can eliminate the ping-pong effect and a lot of useless candidate set. Using users' frequent patterns as well as the probability distribution of time, the user's location can be predicted.

We obtain group's frequent mobility patterns from the individual user. By analyzing the corresponding characteristic group's regularity can be found out. Our work can help operators and ISPs support the location-based services, path planning and forecasting, etc.

Data Set Descrption. The dataset of one city consists of more than 50,000,000 records each day from 17th Nov,2012 to 23th Nov,2012 and the connecting subscribers is 3,270,860 which occupying the 11.1% of the city's population 29,450,000. The data has been collected by a telecom service provider. Each record consists of the users ID, the traffic type history, the web browsing history, the online duration, the start time to connect the base station and the end time to disconnect, as well as the phone numbers, and their LAC, CellID.

II. Apriori Algorithm

Studies have shown that a mobile user often move regularly in the same way .Association rules are rules that describe some potential relationships between data items[9,10]. A typical example of association rules is: In supermarkets, 90% of customers who buy the bread and butter will buy milk. Its intuitive sense is that customers buying a product will tend to buy other goods.

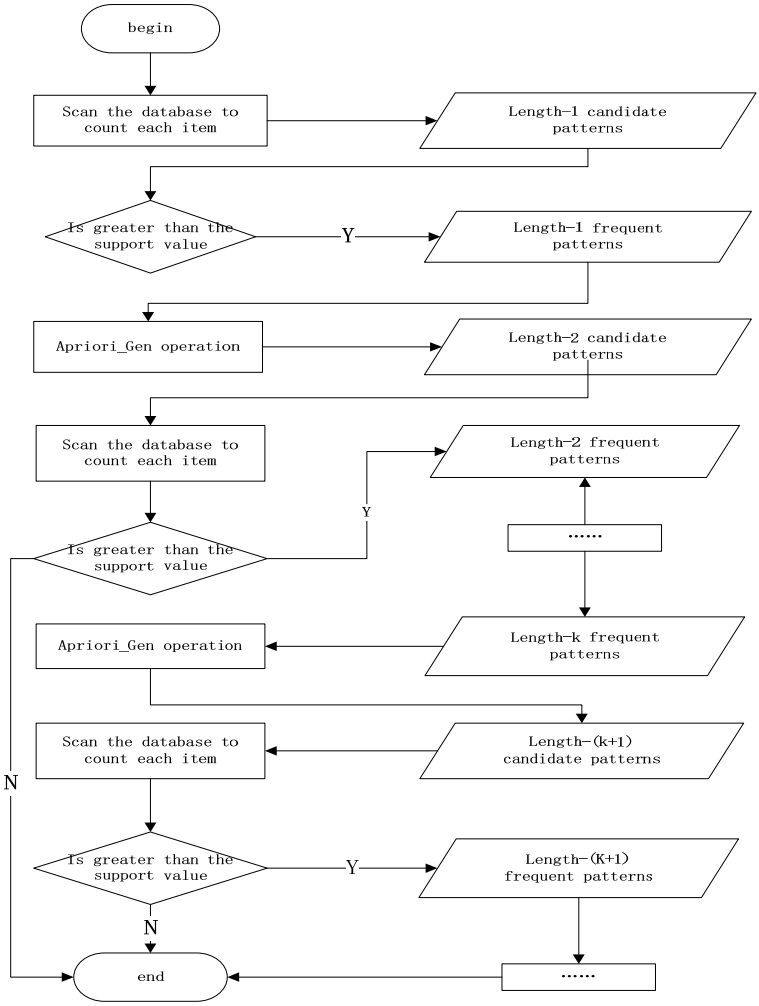

Fig.1 The procedure of Apriori

Association rules are one of the main modes of the current data mining methods focusing on establishing links between the different fields of data. It can find out dependencies between a number of areas under the given condition. Currently the classical algorithm is Apriori, its process is: 1, according to the user's trajectories user mobility patterns are mined; 2, on the basis of these patterns mobility rules are extracted; 3, mobility predictions are accomplished by using these rules

[11]used Apriori algorithm for location prediction in mobile environments. The procedure of the Apriori algorithm about user behavior analysis is as Fig.1.

However, there are two problems in the actual situation.1、Users' records in the base station may sometimes lead to the ping-pong effect, so before finding out users' trajectory of every day we need data preprocessing. 2、Frequent pattern only contains the user's location information, the time information is not taken into account. According to these two problems, we made the following improvement.

A. Data preprocessing. Apriori algorithm uses the hidden information based on historical data to improve the prediction accuracy. If user's daily activities are regular then his location prediction model is a good choice. So we choose the users whose record is greater than 300 per day. As illustrated in Fig.2. These people account for more than forty percent of the total population. By sampling the trajectories of these users it meets the test requirements.

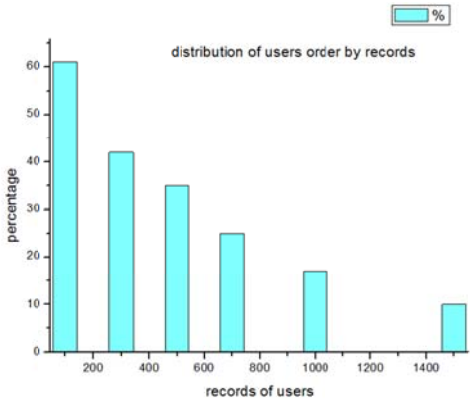

Fig.2 distribution of users order by records

From the records we selected the fields as follows: the phone numbers, start time to connect the base station, the end time to disconnect, CellID the user connect. The records in a day are arranged in chronological order. Thus we get the base station that the user connected. Then we can get the user's trajectories.

But there is a problem in the original data. In urban areas, There are a lot of intersections between base stations. It will lead to the so-called ping-pong phenomena when constructing a user's trajectory. The ping-pong phenomena mean that when the user moves back and forth on the edge of base stations, it will generate transition records in the adjacent base stations. [12]defined the concept of ping-pong phenomena: For Cell i and j, if at any time a user makes a sequence of transitions $i \rightarrow j \rightarrow i \rightarrow j$, then these transitions are ping-pong transitions. Degree of transition is defined: x, y represent two different base stations, using L<x, y> to represent the degree of transition between the two base stations, means the total number of transitions in a period of time. For example, there exists a trajectory Z = [x, y, x, y, x, y, x ...], if threshold value is 3,then you can find <x, y> switching frequency is 5, means that the user produces five records between base station x, y. So the transitions degree is 5.If the transition degree exceeds a given threshold, then the two base stations should be put together .So the trajectory Z change into [x, y]. The simulation results indicate that the proposed algorithm can make accurate handoff decisions and eliminate the ping-pong phenomena compared with the traditional algorithm. In the formula: $T < x, y >$ means a transition of <x，y>, $L < x, y >$ means the degree of transition, r is the records.

$$L < x, y >= \sum_{r=1}^{r=n} T < x, y >. \quad (1)$$

B. The time information in mobility patterns. On weekdays or weekends, there will be obvious difference on the movement path. For example, people usually need to go to work, students go to school on weekdays. While on weekends they may go out for an outing. Meanwhile, within a day, in the morning, afternoon or evening the user's mobile path will have difference too. This is mainly based on this consideration: In the morning users usually go to work or school. In the afternoon the behavior is on the contrary. Moreover, the case of the evening will be more complicated. Therefore, based on the above considerations, the time information of the mobile pattern is divided into two cases to consider:

Case 1: Classified according to the working days and weekends. Each frequent pattern in the frequent movement pattern set stores the distribution of the week(1 represents weekends, 0 represents workdays).

Case 2: The distribution of time. The day is divided into three periods: {[6: 00, 12: 00], [12: 00, 18: 00], [18:00, 6:00]}, and they are represented by 0, 1, 2. The start time of the frequent patterns are divided into three cases, and count the number of instances separately. So the user movement patterns also contain the time distribution of the user's path at all locations It can be analyzed from the distribution that which place is more important for the user.

For example, according to the previous method, then the trajectory of user_A is as follows:

[389, 4247, 389, 7645, 389, 28082, 4247]

If each cell is replaced by a set of <c, x, y>, where c is the cell number, x represents weekend or weekday, y represents the period of the day, then the trajectory of user_A can be expressed as:

[<389,1,0>,<4247,1,0>,<389,1,1>,<7645,1,1>,<389,1,1>,<28082,1,2>,<4247,1,2>]

So when generating frequent patterns, we not only consider the continuous changes between cells, but also consider the change of time, thus frequent patterns are more accurately.

III. Group's Frequent Mobility Patterns

Through the previous work, we constructed individual movement pattern. But, in addition to the analysis of individual, if we are able to dig up some group characteristics from a large number of user mobile patterns, it may be useful in the study of group activities.First we introduce the concept of group support. As shown in the formula: |group| is the total number of users in a group, FP is individual frequent patterns.

$$\text{Group}_{\text{sup}} = \frac{\sum_{i=1}^{i=n} \text{FP appear in user's pattern set}}{|\text{group}|}. \quad (2)$$

To become a frequent pattern of groups, they need to be frequent patterns of one user or some users. If a mobile user mobility patterns is not frequent in the collection of patterns, then they can't be frequent patterns of groups. If a mobility pattern appears in the pattern set in some users, and is greater than the group support threshold, then the mobile pattern should be the movement characteristics of this group.

IV. The Experimental Results and Analysis

We filter out the users whose records are more than 300 every day and analyzed their 7 days ' trajectories. These people account for more than forty percent of the total users. So the frequent patterns and the accuracy of the next location prediction can be found out.

Fig.3 shows the probability distribution of frequent patterns of different length. The horizontal axis represents the length of frequent patterns. The ordinate represents the probability distribution of the frequent patterns. We can find that more than 90 percents of frequent patterns are less than 7, and the highest proportion are length 3 and 4.This can explain that most of the users are active in three or four positions.

Apriori use the trajectory to predict the next position. As we can see from Fig.4, the horizontal axis represents the length of substring for predicting the next position, the ordinate represents the prediction accuracy rate. With the increase of the length of substring, the forecast accuracy has improved. This is mainly because that if a substring is longer, it may have more priori knowledge. Therefore the prediction will be more accurate.

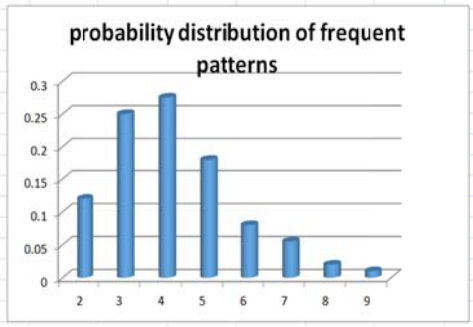

Fig.3 probability distribution of frequent patterns

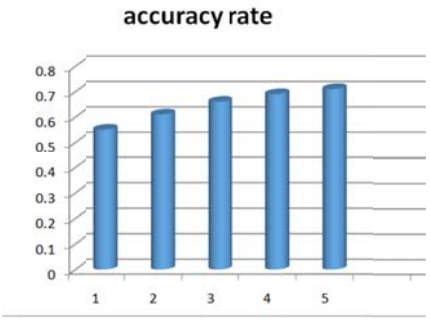

Fig.4 Prediction accuracy of different length of substring

We analyze the probability distribution of frequent patterns on weekdays and weekends to find out whether the frequency of movement patterns of different lengths is the same. As is shown in Fig.4, if a frequent pattern occurs at a very low probability on the weekend, but the probability is very high during the weekdays, to some extent, it may reflect the activity law of the users.

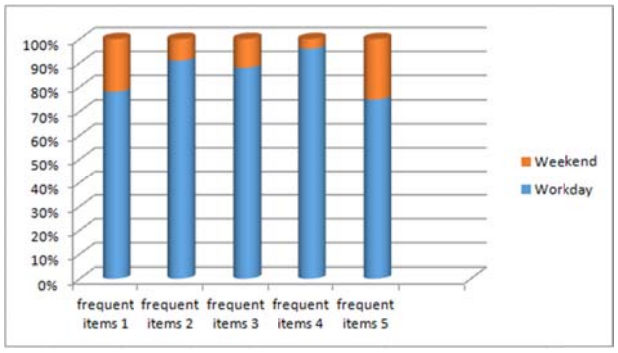

Fig.5 distribution of frequent patterns on weekdays and weekends

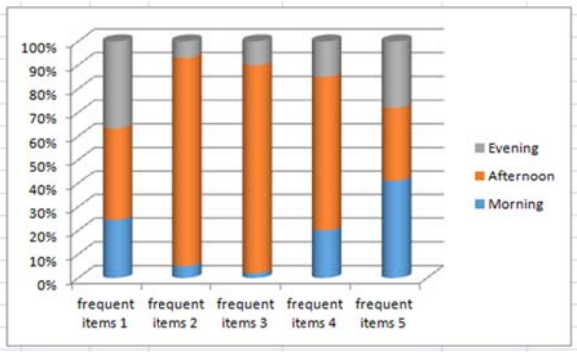

Fig.6 distribution of frequent patterns in a day

In Fig.6, the horizontal axis represents the frequent patterns too. Each frequent pattern corresponding to three time periods, they are morning, afternoon and evening. The ordinate represents the probability distribution of the three time periods. By analyzing the chart we can find that the first four periods of frequent patterns often occur in the afternoon, the fifth session of frequent patterns occurring mostly morning.

Summary

In this paper, we employ Apriori algorithm method to mine mobile users' movement model. It can be concluded that using users' frequent patterns as well as the probability distribution of time,

the user's location can be predicted. Then, we obtain group's frequent mobility patterns from the individual user. By analyzing the corresponding characteristic group's regularity of spatial and time can be found. Our work can help operators and ISPs support the location-based services, path planning and forecasting, etc.

Acknowledgments

This work has been supported by the National Natural Science Foundation of China (61201153), the National 973 Program of China under Grant (2012CB315805), Prospective Research Project on Future Networks in Jiangsu Future Networks Innovation Institute (BY2013095-2-16), the National Basics Research Program 973 of China (2012CB315801), and the Fundamental Research Funds of China for the Central Universities (2013RC0113).

REFERENCES

[1]Pitkänen M, Kärkkäinen T, Ott J, et al. SCAMPI: Service platform for social aware mobile and pervasive computing[J]. ACM SIGCOMM Computer Communication Review, 2012.

[2] Wang D, Pedreschi D, Song C, et al. Human mobility, social ties, and link prediction[C]//Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011: 1100-1108.

[3] Daly E M, Haahr M. Social Network Analysis for Routing in Disconnected Delay-tolerant Manets[C]. In MobiHoc, Florence, Italy , 2007: 32–40.

[4] Zhang Y, Gao W, Cao G, et al. Social-aware data diffusion in delay tolerant manets[M]//Handbook of Optimization in Complex Networks. Springer New York, 2012.

[5] Aschenbruck N, Munjal A, Camp T. Trace-based mobility modeling for multi-hop wireless networks[J]. Computer Communications, 2011, 34(6): 704-714.

[6] Karamshuk D, Boldrini C, Conti M, et al. Human mobility models for opportunistic networks[J]. Communications Magazine, IEEE, 2011, 49(12): 157-165.

[7] Gonzalez M C,Hidalgo C A. Understanding Individual Human Mobility Patterns[J]. Nature, 2009, 453（7196）：779–782.

[8] Borgelt C, Kruse R. Induction of association rules: Apriori implementation[C]//Compstat. Physica-Verlag HD, 2002: 395-400.

[9] Bohannon J. Credit card study blows holes in anonymity[J]. Science, 2015, 347(6221).

[10] de Montjoye Y A, Radaelli L, Singh V K. Unique in the shopping mall: On the reidentifiability of credit card metadata[J]. Science, 2015, 347(6221): 536-539.

[11] Yavaş G, Katsaros D, Ulusoy Ö, et al. A data mining approach for location prediction in mobile environments[J]. Data & Knowledge Engineering, 2005, 54(2): 121-146.

[12] Lee J K, Hou J C. Modeling steady-state and transient behaviors of user mobility: formulation, analysis, and application[C]//Proceedings of the 7th ACM international symposium on Mobile ad hoc networking and computing. ACM, 2006: 85-96.