Mobile Data Mining by Location Dependencies

Jen Ye Goh and David Taniar

Monash University, School of Business Systems, Clayton, Vic 3800, Australia
{Jen.Ye.Goh,David.Taniar}@infotech.monash.edu.au

Abstract. Mobile mining is about finding useful knowledge from the raw data produced by mobile users. The mobile environment consists of a set of static device and mobile device. Previous works in mobile data mining include finding frequency pattern and group pattern. Location dependency was not part of consideration in previous work but it would be meaningful. The proposed method builds a user profile based on past mobile visiting data, filters and to mine association rules. The more frequent the user profiles are updated, the more accurate the rules are. Our performance evaluation shows that as the number of characteristics increases, the number of rules will increase dramatically and therefore, a careful choosing of only the relevant characteristics to ensure acceptable amount of rules.

1 Introduction

Data mining is the process of mining useful knowledge out from a set of raw data. Classical data mining aims to find out knowledge such as association rule [2], sequential pattern [3]. Time series analysis [9-11] can also be applied by using data mining methods so that patterns which are relevant to the decision maker can be found. Mobile data mining [4-7, 12, 13] involves finding out useful knowledge out from mobile users. Outcomes of mobile data mining includes frequency pattern [5], group pattern [13] and parallel pattern [6, 7].

Data mining has now entered a new era of research focusing on analysing the event sequences that is happening in a mobile environment. This is known as *mobile data mining* [5-7]. The essence of mobile data mining consists of a sequence of events happening over a time series, along with the ability of mobile units to be able to move within the coverage area. Processing power and memory capacity of mobile equipment is expensive and could occasionally be down due to poor receptions.

This paper describes the process of location based data mining in the mobile environment by using user profile method. There are different kinds of mobile devices in the mobile environment. These include the mobile phone, Personal Digital Assistant (PDA), laptop and car. Some of these mobile devices have the ability to reply a signal back to the static station and some mobile device either cannot or is too expensive to reply a signal back.

Z.R. Yang et al. (Eds.): IDEAL 2004, LNCS 3177, pp. 225–231, 2004.

2 Background

The mobile environment described in this paper consists of a set of mobile devices, which can be mobile phone, PDA and many more. The static devices described in this paper are devices such as the wireless access point, which stays static in a particular location over time, which mainly have two purposes. First, it is to provide resources to the mobile devices such as bandwidth. Second, it is to record the user visiting data by identifying a particular mobile device in the mobile environment. The static device will regularly send the list of mobile users to a central location, which contains the user profile, and update the user profile.

The concept of location dependency described in this paper means the relationship of the knowledge about mobile user produced which are closely associated with the particular location. Location dependent knowledge is useful for decision making in relation to a particular location. Often, a decision unit is limited to a particular location and therefore, location dependency will support the decision unit.

A related work in mobile mining is group pattern mining [13]. Group pattern [13] aims to find out a set of group, which are nearby to each other over a distance and time. A frequency pattern is produced from the mining of raw data of mobile user based on their physical distance data, and time series data. A group of mobile users can be qualified as a group pattern when they meet the criteria of both physically close to each other below a certain distance threshold, and being physically close to each other over a certain time threshold [13].

3 Proposed Method

3.1 Initial Requirements

In a mobile environment, it consists of a set of mobile devices and also a set of static devices, which tends to be the access points for the mobile devices providing bandwidth for communication and authentication for mobile network access. Each of the static devices have a set of characteristics, ($c1$, $c2$, …, c_n). These characteristics can be generic characteristics such as entertainment, sports, education, shopping to more detailed characteristics such as *comedy.entertainment, badminton.sports, law.education* arranged by means of hierarchical organisation.

3.2 Raw Mobile Data Collection

As mobile devices moves along these static devices, the mobile devices are configured in such a way that it will transmit a identification signal at regular time interval which can be the hardware address and is unique worldwide, thus the ability to uniquely identify a particular mobile device. At the static device end, these signals

are received and recorded. At the end of the process, the raw data collected from the static devices will consist of the following format: (*device_id, static_device_id*).

The *device_id* is the unique identification mark of the mobile device. The *static_device_id* is the unique identifier for the static device. The raw data from each static device can then be gathered and be represented by the following format {*device_id, static_device_id[*($c1$, $c2$, …, c_n)]}. As the mobile users moves along the static devices, user profiles are generated at the same time.

3.3 User Profile Updating Process

User profile consists of a set of characteristics, which the users enjoy. The set of characteristics are found from the static device, as the characteristics are pre-recorded. The more the mobile users visits a particular static device, the set of characteristics listed in the static device will be updated to the user profile more often.

Each mobile user characteristics have a value of 0 % to 100%. The higher the percentage, the higher the indication of the mobile user has visited the static device recently and frequently. Mobile user identification and the list of characteristics represent a user profile. A list of user profiles that visited a particular location is then extracted, and passed to the data mining system to find out association rules from the list. The result of this process is a list of association rules with *support%* and *confidence%* that if *characteristic*1 and *characteristic*2 is higher than the *characteristic_threshold* in one of the user profile, *characteristic*3 will also be higher than the *characteristic_threshold* in the same user profile. A user profile with a characteristic higher than the *characteristic_threshold* would mean that the particular user visits locations, which contain those characteristics.

3.4 Algorithms for Proposed Method

The algorithm for assigning characteristics is described as below. Each location is equipped with a static device, which can be used to communicate with mobile device. Each static device is assigned to a list of characteristics that represents the overall theme of the location, such as {*trainstation.transport*, *cinema.entertainment*, *grocery.shopping*} may represent a train station, with a cinema and grocery shopping center nearby the wireless coverage area.

Figure 1 provides the algorithm and result of user profiling. The *VisitedLocation* contains a list of characteristics that a particular location contains, such as a list of {*badminton.sports*, *comedy.entertainment*}. The identified characteristics are then added into the mobile user profile, with each characteristics an assigned percentage value, such as {*badminton.sports*=0.5, *comedy.entertainment*=0.3} and if the location contains comedy.entertainment, the list will become {*badminton.sports*=0.5, *comedy.entertainment*=0.33}.

```
Function Train User Profile (MobileUser, VisitedLocation) {
    VisitedLocation = V;
    MobileUser.Update (V.Char1, V.Char2, …, V.Char3)
    # MobileUser.Char1 = MobileUser.Char1 + (MobileUser.Char1 * V.Char1/ 10)
}
User Profile Database {
   User A = {comedy.entertainment=0.7, badminton.sports=0.8, law.education=0.5}
   User B = {comedy.entertainment=0.4, badminton.sports=0.3, law.education=0.9}
   User C = {trainstation.transport=0.8, grocery.shopping=0.5, drama.entertainment=0.4}
   User D = {comedy.entertainment=0.6, badminton.sports=0.5, infotech.education=0.4}
}
```

Fig. 1. Algorithm to Train User Profile & User Profile.

As the visiting location of the mobile user data are being collected, they are used to train the user profile database to better represent the overall life picture of the mobile user. The static device then starts to collect the identification code of the mobile device over a certain period of time. The result of this data collection would be: *Location1* = {*User A*, *User C*, *User D*}. The algorithm to retrieve the user profile record is described as below.

The list of user profile retrieved is then passed to a mining algorithm, such as association rule mining algorithm. In the above example, the user profiles retrieved are as below. The *confidence* value for both *comedy.entertainment* and *badminton.sports* coexist in the same transaction is = 2/3 = 66%. Considering the *threshold* value is 55%, this association rule exists. Therefore, the conclusion for this mining exercise is that, the current location has a *cinema.entertainment* background. Although most mobile users visited this location has a *cinema.entertainment* profile, it is found that mobile users who visited this particular physical location not only likes the location theme but also likes *comedy.entertainment* and *badminton.sports* at the same time.

```
Function Retrieve Profile (UserList) {
   Return UserList.1, UserList.2, UserList.3, …, UserList.N;
}
User A = {comedy.entertainment=0.7, badminton.sports=0.8, law.education=0.5}
User C = {trainstation.transport=0.8, grocery.shopping=0.5, drama.entertainment=0.4}
User D = {comedy.entertainment=0.6, badminton.sports=0.5, infotech.education=0.4}
```

Fig. 2. Algorithm to Retrieve User Profile & User Profile Structure.

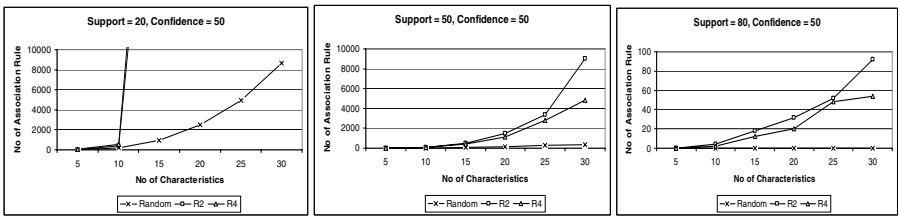
Figure 2 is useful for the decision makers to make a more informed decision by having the knowledge of the association of interests of the mobile users that visited a particular physical location, which was found based on the overall life picture of the mobile user.

4 Performance Evaluation

The performance evaluation was tested on a Pentium IV machine, equipped with 384MB of RAM. The association rule mining [2] process for the performance testing lasts from 1 second to 7 seconds. Three sets of data are generated. The first set is random data, which has been generated from random.org [8]. The random data set is random in terms of the display of random integer of 1 or 0 based on atmospheric noise. The source data consists of 200 records of mobile users visiting a particular location. The association rule mining software is XLMiner Demo Version [1].

The set of random data is labelled as Random. An integer of 1 will represent that the user characteristics has reached greater or equal to the acceptable threshold, say, 60%. Every single piece of data in Random has equal chance of occurring. The other set of data, R2, is produced from Random. R2 aims to show the repetition characteristics of mobile users and aims to produce more repetitions of similar user characteristics in the list. In R2, the first two records are repeated every next eight mobile users.

The dataset R4, which have the concept similar to R4, is obtained from Random with the first four record repeated every six mobile users. This shows a much more repetition of similar user characteristics for mobile users visiting a particular mobile location. There are instances when the number of rules is too high and the system refused to output to prevent crashes.



(a) (b) (c)

Fig. 3. Performance Chart Using Various Supports.

Figure 3(a) above has shown a gradual steady increase in the number of association rules [2] as the number of user characteristics increases for the Random dataset. However, the number of association rules becomes too many to the extent that the data mining software have rejected the mining process at 15 user characteristics level for dataset R2 and R4. Figure 3(b) have shown that with the *support*=50, *confidence*=50, the number of rules generated from Random, R2 and R4 increases gradually. There are more rules found in R2 than R4, and there are more rules found in R4 than Random. All dataset have a gradual increase over the number of user characteristics. Figure 3(c) have shown that with the *support*=80, *confidence*=50, the number of rules generated from R2 and R4 are non linear. For Random, due to the nature of random data which every single number have equal chance to occur, no rules are generated with stronger support threshold. From the graph, it can be seen that R4

always have lesser rules found than R2. At the point of 25 user characteristics, the magnitude of difference is relatively lesser than other readings. But overall, R2 and R4 increase over the number of user characteristics. Figure 3(c) also have the y-axis range from 0 to 100 because the range of number of rules significantly reduces as the support threshold is increased, thus only rules with very high confidence are presented.

The conclusion is increase of user characteristics leads to significant increase in the number of association rules [2] mined. In Figure 3(c), Random set increased in a steady fashion, but R2 and R4 have generated too much rules by 15 characteristics. This suggests that when mining is performed ensure to choose only the relevant set of characteristics for quicker and more relevant rule generation.

5 Conclusion and Future Work

User profile being represented as a set of characteristics and percentage value represents how much the user is likely to be involved with a particular characteristic, based on past information. The more frequent the user profile is updated, the more meaningful the user profile. It was found that as the number of characteristics increases, the number of rules found increased significantly. Therefore, careful choosing of the set of characteristics of user profile should be done before mining the source data in order to improve performance and economy.

Future work is to find out time dependent knowledge of location based knowledge. Time dependent knowledge involves putting a timestamp for each knowledge found and gives an expiry date for each knowledge.

References

1. XL Miner. Cytel Software Corporation, 2004.
2. R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules. In Proc. 20th Int. Conf. Very Large Data Bases, pp. 487-499, 1994.
3. R. Agrawal and R. Srikant. Mining Sequential Patterns. In Proc. 11th Int. Conf. on Data Engineering, pp. 3-14, 1995.
4. J. Goh and D. Taniar. Mining Density Pattern from Mobile Users. 2004. (submitted)
5. J. Goh and D. Taniar. Mining Frequency Pattern from Mobile Users. Knowledge-Based Intelligent Information & Eng. Sys., 2004. (accepted)
6. J. Goh and D. Taniar. Mining Logical Parallel Pattern from Mobile Users. Int. Conf. on Intelligence in Communication Systems, 2004. (submitted)
7. J. Goh and D. Taniar. Mining Parallel Pattern from Mobile Users. Int. Conf. on Embedded and Ubiquitous Computing, 2004. (submitted)
8. M. Haahr. True Random Number Service. Random.org, 1998.
9. J. Han, G. Dong, and Y. Yin. Efficient Mining of Partial Periodic Patterns in Time Series Database. In Proc. of Int. Conf. on Data Engineering, pp. 106-115, 1999.

10. J. Han, W. Gong, and Y. Yin. Mining Segment-Wise Periodic Patterns in Time Related Databases. In Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining, vol. no. pp. 214-218, 1998.
11. J. Han, J. Pei, and Y. Yin. Mining Frequent Patterns without Candidate Generation. In Proc. Int. Conf. SIGMOD, pp. 1-12, 2000.
12. E.-P. Lim, Y. Wang, K.-L. Ong, and et al. In Search of Knowledge About Mobile Users. ERCIM News, vol. 1, no. 54, pp. 10, 2003.
13. Y. Wang, E.-P. Lim, and S.-Y. Hwang. On Mining Group Patterns of Mobile Users. In Proc. of DEXA, pp. 287-296, 2003.