Mining Physical Parallel Pattern from Mobile Users

John 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 Data Mining focuses on finding useful knowledge out from raw data from mobile users. In this paper, parallel pattern is proposed, which describes the movement trend patterns of mobile users. Parallel pattern aims to find out the trend of movement based on physical location as detected by the wireless access point. The performance testing of this paper shows that as the number of different mobile user increases, under random distribution, the number of parallel pattern found decreases. Therefore, it is important to adjust the size of the window depending on the amount of mobile users surrounding the particular wireless network access point.

1 Introduction

Mobile Data Mining [5, 6, 12] is a field under data mining which focuses on finding useful knowledge from raw data from mobile devices. Mobile devices can consists of mobile phone, personal digital assistant, laptop computers, cars and many more. As long as a device can move and can be identified, then there is an avenue for mobile data mining.

The inspiration for mobile data mining is that mobile devices often follow the owner most of the time. Mobile phone is a good example. It often stays close to the owner and registers the presence on the mobile phone network. By performing mobile data mining on these sets of raw data, knowledge found will better represent the overall lifestyle of the user. Comparing this with classical data mining, mobile mining have the advantage to take the big picture of the owner lifestyle, at all times and in all locations. Classical data mining [1, 2, 11], on the other hand often focus on a particular set of function, such as market basket data analysis, which focuses on specific area such as purchasing habit.

Other fields in data mining includes the time series analysis [8, 9], mining frequent patterns [10], web pattern mining [3, 4, 15], and spatial data mining [11]. Our previous work in mobile data mining consists of [5, 6]. This proposed method is a part of our mobile data mining project in search of finding useful methods and algorithms for mobile data mining. Another paper based on parallel pattern on logical preference characteristics of mobile users will be published soon.

Mobile data mining is still in the beginning stage of research and requires significant amount of contribution. With the time series data of mobile users travelling around the mobile coverage area with each area labelled with different characteristics, many interesting knowledge can be found. Some of the proposed method for mining

L.T. Yang et al. (Eds.): EUC 2004, LNCS 3207, pp. 324–332, 2004.

useful knowledge from mobile users includes group pattern, frequency pattern and location dependent characteristics pattern.

In a mobile environment, the static nodes are the physical equipments that provide resources to the mobile users. An example of static node is the wireless network access point where mobile equipments will contact the static node to register their presence within the coverage area and use the bandwidth resources. The mobile nodes are the physical mobile equipments that move around the mobile environment. The mobile nodes will request services from static nodes, by revealing their identification code. It is assumed that mobile node will have limited amount of resources, such as memory, storage capacity and processing capability. It is also assumed that static node will have sufficient level of memory, storage capacity and processing power. [5, 6]

The remaining of this paper is sequenced as below. Section 2 describes the related work to mobile data mining which consists of group pattern, frequency pattern and location dependent data mining. Section 3 describes the proposed method, which is the parallel pattern. Section 3 consists of step by step approach on describing how to mine parallel pattern with the algorithms. Section 4 describes the performance evaluation of the proposed method. Finally, Section 5 summarises the paper and draw conclusions based on the proposed method.

2 Related Work

There are a number of related works in the field of mobile data mining. These include the group pattern [13, 14] which uses physical distance to find out groups of users. The frequency pattern is developed in order to address the inherent problem of group pattern that is mobile equipment users often use their mobile equipment when they are far away from each other. Therefore, the frequency pattern [5] uses the logical distance as a means to find out groups of users. The location depending mobile data mining describes a method in which location dependent knowledge can be found by profiling the users and performing data mining on their user profiles that visits a particular location. [5]

2.1 Group Pattern

Group pattern [13, 14] was proposed by some researchers in Singapore. A group pattern is represented by a number of mobile user identification. The group pattern tells the high possibility of members of the group is closely related group of people. Members in a group pattern have to be close to each other in terms of physical distance and be close to each other over a certain timeframe. The physical distance is calculated by using Euclidean distance.

The weakness of group pattern is the inability to use logical distance instead of physical distance. [5] Logical distance makes more sense in a mobile environment when mobile coverage is large and mobile users that are close to each other physically do not use their mobile device for communication. Mobile users rather use mobile devices to communicate with closely related persons that are located far away in terms of physical distance. In order to address this issue, frequency pattern was proposed.

2.2 Frequency Pattern

Frequency pattern [5] was proposed by us to address the inherent problem of group pattern of using physical distance as the measurement. Frequency pattern proposed to use logical distance that is calculated by using frequency of communication as the measurement to qualify as a group. As mobile users communicate with each other more often using mobile devices over a geographical distance, it indicates that there exists a certain relationship between the entities involved in the communication process.

Frequency pattern further enhance the calculation by the ability to give different emphasis on different parts of the timeframe so that more recent communications can be treated as more important than less recent communications, thus enhancing the calculation of relative frequency between two mobile users. [5]

2.3 Location Dependent Mobile Data Mining

Location dependent mobile data mining [6], another piece of our work, proposes a way for finding useful knowledge regarding the taste of the mobile users by means of assigning characteristics to each static node in the mobile environment. Each mobile user will be registered with the theme of the location it visited and will be shown as the list of characteristics that a particular mobile user has interest. A high frequency of characteristics count will represent a strong interest in a particular theme, such as *shopping* or *library*.

The list of characteristics that exceeds the strong interest threshold is then passed to association rule mining algorithm to find out the association rule of the characteristics of mobile users in one particular location. The result of this data mining process is a set of associative relationship between the characteristics of mobile users. One such example is that there is 80% of confidence and 50% of support that mobile users that visited location A have strong interest in both *library* and *comedy*. [6]

3 Proposed Method: Parallel Pattern

Parallel pattern is essentially the high correlation of two actions happening at the same time (m_1, $s_1 \rightarrow s_2$ & m_2, $s_2 \rightarrow s_3$). Parallel pattern is a newly proposed terminology. A mobile environment consists of multiple mobile devices (m_1, m_2, …, m_n) moving around static devices (s_1, s_2, …, s_n) that provide resources to the mobile devices. Each static device (s_n) consists of a few characteristics (c_1, c_2, …, c_n) that represents the overall theme of the location of the static device. An action in a mobile environment involves a mobile device moving from one location to another (m_1, $s_1 \rightarrow s_2$). As two or more actions must happens together quick enough to present the overall parallel effect, therefore, only actions that happens within the first 5 seconds and the next 5 seconds are taking into the calculation.

The definition of an action is the occurrence when a mobile node moves from one static node to another static node. Each action consists of mobile node identification (m_n), movement from static node (s_i), movement to static node (s_j). Each action is recorded by means of information sharing between static nodes. Once each static node is recorded, calculation of parallel pattern for mobile data mining can commence.

3.1 Finding Parallel Pattern

The following represents a list of actions that was recorded for past 10 seconds. The format of an action is recorded as (*mobile node identification*, *source static node* → *destination static node*). The time series is represented as (t_1, t_2, …, t_n). The following represent a sample dataset of the purpose of describing the mobile data mining process.

t_1: (m_1, $s_1 \rightarrow s_2$)
t_2: (m_2, $s_1 \rightarrow s_2$)
t_3: (m_3, $s_1 \rightarrow s_2$)
t_4: (m_4, $s_1 \rightarrow s_2$)
t_5: (m_1, $s_2 \rightarrow s_3$)
t_6: (m_2, $s_2 \rightarrow s_3$)
t_7: (m_3, $s_2 \rightarrow s_3$)
t_8: (m_4, $s_2 \rightarrow s_3$)
t_9: (m_1, $s_3 \rightarrow s_4$)
t_{10}: (m_2, $s_3 \rightarrow s_4$)

Fig. 1. Sample Time Series Raw Data

Figure 1 consists of m_1, m_2, m_3 and m_4 moving the same direction one after another. As long as the actions occurred is within the pre-specified window size, in this case is 10 seconds, they are considered parallel.

Step 1: Calculating Frequency

Frequency = No. of Same Movement Pattern / Total Number of Movements. The frequency represents how frequent the occurrence of similar movement pattern. For the movement pattern of ($s_1 \rightarrow s_2$), it happened in (t_1, t_2, t_3, t_4). Therefore, frequency is calculated as: 4 / 10 = 40%. For the movement pattern of ($s_2 \rightarrow s_3$), it happened in (t_5, t_6, t_7, t_8). Therefore, frequency is calculated as: 4 / 10 = 40%. For the movement pattern of ($s_3 \rightarrow s_4$), it happened in (t_9, t_{10}). Therefore, frequency is calculated as: 2 / 10 = 20%.

Step 2: Discard Patterns Lower Than Frequency Threshold

The purpose of discarding patterns lower than frequency threshold is that those lower than the threshold level is unlikely to be important for decision making purposes, and best to not taken into consideration. In the above example, with a threshold of 40%, movement pattern 1 and 2 are accepted while movement pattern 3 is rejected. The list of movement pattern are ($s_1 \rightarrow s_2$) and ($s_2 \rightarrow s_3$).

Step 3: Calculating Confidence

Confidence = No. of Same Movement Pattern / Total Remaining Number of Movements. In the above example, actions that do not meet the frequent threshold requirement are discarded. Therefore, the total number of actions goes from 10 to 8. The confidence of each action is calculated. In the above example, confidence of ($s_1 \rightarrow s_2$) is 4 / 8 = 50% and confidence of ($s_2 \rightarrow s_3$) is 4 / 8 = 50%.

Step 4: Generation of Parallel Pattern

Parallel Pattern is a list of similar movement patterns that occurs significantly frequency enough (greater or equal to frequency threshold) and have significant confidence (greater or equal to confidence threshold). The patterns that satisfy the requirement will then be listed in the system output.

Each parallel pattern will be represents with a movement pattern along with the frequency and confidence in the following format: ($s_1 \rightarrow s_2$, 40%, 50%), ($s_2 \rightarrow s_3$, 40%, 50%). The parallel pattern describes the parallel movement of similar patterns from the raw data source. In this case, the similar pattern is the movement are such as movement from s_1 to s_2 and movement from s_2 to s_3.

The parallel pattern found describes how each unit in the mobile environment acts in parallel. The mobile nodes that move together in parallel present a set of group that are related to each other in term of movement decision making. It may consist of a leader and many followers that follow what the particular leader does.

3.2 Algorithm

Figure 2 represents the algorithms required to calculate parallel pattern. The function Calculate Frequency calculates the frequency for all actions by dividing the pattern register with the window size. Then, the function Discard Window checks each action to ensure that actions that have a frequency below the frequency threshold will be discarded. Then, the function Calculate Confidence calculates the confidence of the actions that are frequent. It is done by dividing the pattern register with the new window size, which is original window size less the discard of infrequent actions. Finally, the function Generate Parallel Pattern generates the parallel pattern by listing out all actions with frequency and confidence above the frequency threshold and confidence threshold.

The algorithms consist of multiple functions and a main body structure. The parallel pattern is generated by first finding all the frequency of each action and discards all actions that are infrequent as they are not significant to be considered as a pattern. After the discard process, the confidence for each set of pattern is calculated based on the percentage of occurrence after the discard process. This way, a stronger parallel pattern can be assured. Finally the parallel pattern generation process will generate each unique action that satisfies the two constraints that is frequency greater than frequency threshold and confidence greater than confidence threshold.

```
 Function Main {
Calculate Frequency;
Discard Window;
Calculate Confidence;
Generate Parallel Pattern;
 }

 Function Calculate Frequency {
For I = 1 to Window Size Do
 Increment Pattern Register by 1;
End For
Frequency = Pattern Register / Window Size;
Return Frequency;
 }

 Function Discard Window (Frequency Threshold) {
For I = 1 to Window Size Do
 If Current Action.Frequency ≤ Frequency Threshold Then
 Delete Current Action;
 End If
End For
 }

 Function Calculate Confidence {
For I = 1 to (Window Size – Total Discard) Do
 Increment Pattern Register by 1;
End For
Pattern.Confidence = Pattern Register / (Window Size – Total Discard);
Return Pattern.Confidence;
 }

 Function Generate Parallel Pattern {
For I = 1 to (Window Size – Total Discard) Do
 If (Action.Frequency ≥ Frequency Threshold) AND
 (Action.Confidence ≥ Confidence Threshold) Then
 Display Action, Frequency, Confidence;
 End If
End For
 }
```

Fig. 2. Algorithms for Mining Parallel Pattern

4 Performance Evaluation

Performance testing was done under a Pentium IV computer equipped with 384MB of RAM and 1GB of free hard disk space. The main aim of this performance evaluation test is to find how accurately the proposed method can find parallel pattern given various different known raw sets of raw data from the mobile environment. The measurement to determine how accurately the proposed method has found the knowledge is by measuring the number of parallel patterns found.

The raw data consists of a set of randomly generated data based on atmospheric noise [7]. Dataset A, Dataset B and Dataset C are a set of randomly generated numbers which contains 10, 20 and 30 unique actions respectively. The aim of the performance testing is to find out the relationship between different amounts of unique action, window size towards the number of parallel pattern that can be found.

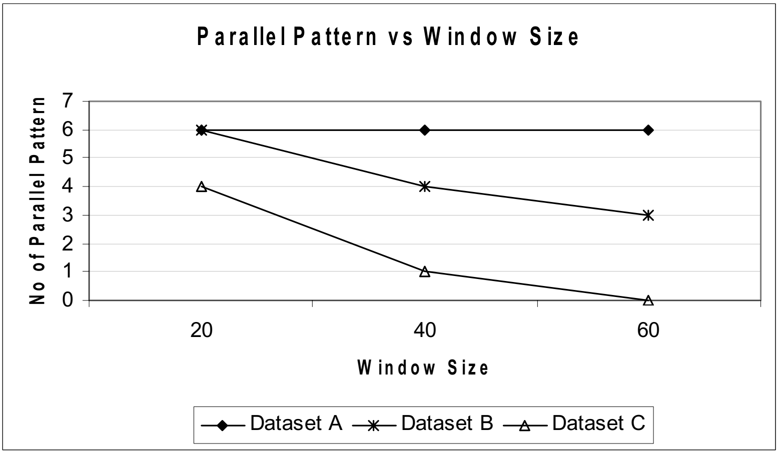

Fig. 3. No of Parallel Pattern vs Window Size

Figure 3 represents the result of the performance testing. Three data sets are supplied and they are called Dataset A, Dataset B, and Dataset C. Dataset A contains 10 different unique actions. Dataset B contains 20 different unique actions. Dataset C contains 30 different unique actions. By using these three datasets to test against different window sizes, the number of rules found is the output of the test. Each integer represents a unique action identification number.

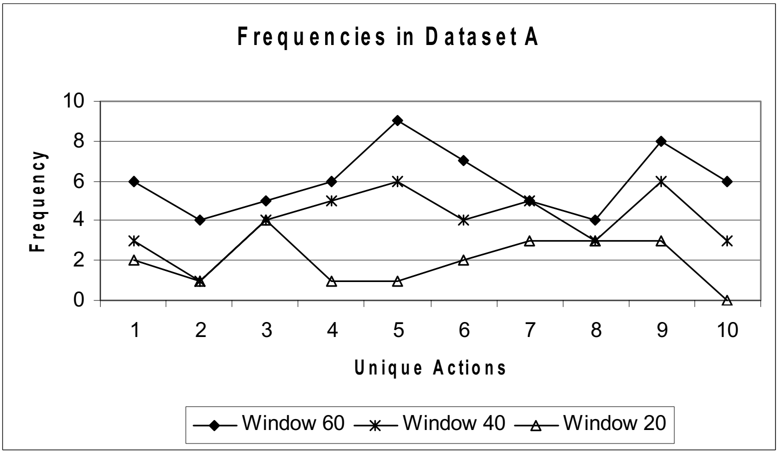

Fig. 4. Frequencies in Dataset A

Figure 4 shows the frequencies distribution for Dataset A. The graph serves as an indication of the randomness of the source data. One interesting observation that can be found this graph is that the larger the window, the higher the frequency of the particular unique action. Therefore, when dealing with the real life dataset, it is important to adjust the window size accordingly due to the sensitivity demand. Generally, larger window size yields more patterns and smaller window size yields lesser patterns.

The result shows that as the window size increases, Dataset C always have lesser parallel pattern than Dataset B, and Dataset B always have lesser parallel pattern than Dataset A. Therefore, the number of parallel pattern reduces as both the window size increases and when the number of unique actions increases. It is also interesting to note that the number of parallel pattern found for Dataset A stays at 6 with increase of window size. Due to the number of unique actions is small, 10, the frequency of occurrence is always high.

5 Conclusion and Future Work

Parallel pattern truly represents an important knowledge to be found from the raw data collected from the mobile device. It provides the support for decision makers to make decisions which involve the need for the knowledge of surface movement pattern of mobile users. Armed with this piece of knowledge, by mapping the knowledge to the current two dimensional physical map of the mobile environment, interesting patterns can be seen.

The future work for this research is by incorporating the parallel pattern with location dependency mobile data mining, a higher level of knowledge can be found. Rather than describing how mobile nodes moves in parallel, it has the power to describe how mobile nodes seeks different sets of characteristics based on the parallel logical movement rather than parallel physical movement that is described in this paper.

References

1. R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules. *Proc. 20th Int. Conf. Very Large Data Bases*, pp. 487-499, 1994.
2. R. Agrawal and R. Srikant. Mining Sequential Patterns. In *Eleventh International Conference on Data Engineering*, 1995.
3. V. Christophides, G. Karvounarakis, and D. Plexousakis. Optimizing taxanomic semantic web queries using labeling schemes. *Journal of Web Semantics*, vol. 1, pp. 207-228, 2003.
4. M. Eirinaki and M. Vazirgaiannis. Web Mining for Web Personalization. *ACM Transactions on Internet Technology*, vol. 3, pp. 1-27, 2003.
5. J. Y. Goh and D. Taniar. Mining Frequency Pattern From Mobile Users. *In Proc. KES* 2004. (To Appear)
6. J. Y. Goh and D. Taniar. Mobile Mining By Location Dependencies. *In Proc. IDEAL 2004.* (To Appear)
7. M. Haahr. Random.org - True Random Number Service. 1998.
8. J. Han, G. Dong, and Y. Yin. Efficient mining of partial periodic patterns in time series database. *ICDE*, pp. 106-115, 1999.

9. J. Han, W. Gong, and Y. Yin. Mining Segment-Wise Periodic Patterns in Time Related Databases. *4th International Conference on Knowledge Discovery and Data Mining*, pp. 214-218, 1998.
10. J. Han, J. Pei, and Y. Yin. Mining Frequent Patterns without Candidate Generation. *In Proceedings of International Conference SIGMOD 2000*, vol. 24, pp. 1-12, 2000.
11. K. Koperski and J. Han. Discovery of Spatial Association Rules in Geographical Information Databases. *4th International Symposium on Advances in Spatial Databases*, pp. 47-66, 1995.
12. D. L. Lee, J. Xu, B. Zheng, and W.-C. Lee. Data management in location-dependent information services. *Pervasive Computing, IEEE*, vol. 1, pp. 65-72, 2002.
13. E.-P. Lim, Y. Wang, K.-L. Ong, and et al. In Search Of Knowledge About Mobile Users. *ERCIM News*, vol. 54, 2003.
14. Y. Wang, E.-P. Lim, and S.-Y. Hwang. On Mining Group Patterns of Mobile Users. *Lecture Notes in Computer Science, DEXA 2003*, vol. 2736, pp. 287-296, 2003.
15. Y. Xiao and J. F. Yao. Traversal Pattern Mining in Web Usage Data. Chapter from Web Information Systems. 2004.