# Mining Frequency 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.** Group pattern was introduced to find groups of mobile users associated by means of physical distance and amount of time spent together. This paper addresses the inherent problem of group pattern, that mobile user are often not physically close together when they use mobile technology, by proposing frequency pattern. Frequency pattern use creative method to calculate frequency of communication between mobile users. By using frequency rather than physical distance, the closeness of two mobile users can better be represented. Performance of the proposed method indicates a suitable segment size and alpha value needs to be selected to get the best result.

## 1 Introduction

Data Mining is an area of research that focuses on finding out useful knowledge from a set of data sources [1]. The data sources can come from time series data, spatial based data, transaction based data and many more. The goal is find useful knowledge suitable for decision makers. Within data mining, a branch of mobile data mining can be extended based on the special conditions of mobile environment. In a mobile environment, there are nodes, which can be cars, mobile phones, PDAs. These nodes are often equipped with limited resources, such as limited battery life, memory capacity and processing capacity [2].

Group pattern [4] represents a group of users which stays within a maximum level of distance magnitude over a minimum amount of time. For example, a group of mobile nodes which is together over a short distance and over a specified time would be considered as a group. Therefore, a group pattern represents a group of mobile users which are related together in terms of distance and time. [4]

Group pattern has its unique characteristics independent from association rules [1], as association rules are transaction based while group pattern is based on dynamic time series. Group pattern is also distinct from clustering [3] because clustering is based on grouping entities that have similar characteristics and features, while group pattern is based on the movement data of mobile users which are dynamic as the time and distance changes frequently. Group pattern uses Euclidean distance [4] as a method to determine whether to include a set of mobile nodes into a group, forming a group pattern. Unfortunately, the purpose of mobile applications is to perform tasks without distance limitation.

M.Gh. Negoita et al. (Eds.): KES 2004, LNAI 3215, pp. 795–801, 2004.

## 2 Related Work

There are relatively few papers written specifically for methods in finding useful knowing in the mobile environment, also known as mobile mining. One of the related work that aims to finding useful knowing of groups of mobile users was presented in. [4] The input contains a database of raw data which consists of mobile user identification, and their two dimensional location data over a time series. The process of the data mining involves determining groups of users that stays together over two variables (time and distance).

The aim of this process is to provide a list of groups of mobile users, which during the time series has a magnitude of distance lesser than the distance threshold, and has a magnitude of time greater than the time threshold. In another words, if there are a few mobile users that are close to each other over a certain time, it is then considered as a group pattern [4].

The advantage of group pattern is that it is only suitable in finding mobile user knowledge in an environment that is restricted, that is, mobile users moves within a closely monitored region. This is because of the process involves finding out the $x$ and $y$ axis of the location of the mobile user data, which can be very expensive in real life.

The disadvantage of group pattern is that, there is an inherent problem of the entire process that is the method was used in mobile environment. A mobile environment is an environment that mobile nodes move freely over a large geographical distance. The purpose of mobile equipment usage is to break the barrier of distance, enabling users to interact with others without needing to get physically close to each other.

Group pattern is only useful when mobile users are actually close to each other, as a group is a group of mobile user that is close to each other over time. This, in real life can rarely happens, as if two people are physically close to each other, they don't need to use mobile devices to communicate to each other. Furthermore, in real life, it may return a lot of noises (wrong result), especially mobile nodes are physically close to each other in public places, such as bus stops, airport terminals, while they are waiting for something to happen. The above example can easily qualified as a group pattern, as a group of mobile users are close to each other over a certain time. But these mobile users may not really have much interaction with each other when they are close physically. Rather they might be busy sending communications using their mobile equipment to people in another country.

## 3 Frequency Pattern: Proposed Method

In the area of data mining, there is mobile mining which focuses on the goal of finding out useful knowing based on raw data of mobile users. In the area of mobile mining, the proposed method is called frequency pattern meets the goal of mobile mining of finding useful knowledge from raw data of mobile users by means of flexible method for calculating a frequency of communication between two mobile nodes and determining their logical proximity and returns useful knowledge of sets of mobile users that are close to each other.

Frequency pattern is discovered by calculating the relative frequency between two mobile nodes. First, the pre-specified criteria, such as the number of segments, the *alpha* values and the size of the time series to be considered are provided. Then, a relative frequency is calculated. In order to determine whether the relative frequency is strong enough to be generated as knowledge, a threshold has to be set. If the relative frequency is stronger than the threshold, then it will be accepted. The final outcome is a knowledge, which contains mobile nodes and their strong relative relationships among each others.

Let the time series in consideration be represented as {*t*1, *t*2, …, *t*10} where each item represents an unit of timeslot. Let the segments be represented in the manner of segment1{*t*1 *to t*3}, $\alpha$=0.2, which represents segment number 1 ranges from timeslot *t*1 to timeslot *t*3 inclusive, with an *alpha* value of 0.2. The addition of all *alpha* values in all segments must equate to 1. In another words, for the final relative frequency value, the *alpha* value for each segment represents the amount of emphasis is given for a particular timeslots in a particular segment. All the conditions mentioned are called the *pre-specified criteria*. The pre-specified criteria is important because it enables the frequency calculation to be tailored to different data mining situations by providing different emphasis on different segments of the time series.

The calculation of relative frequency between two mobile nodes is non-directional. This means that a 0.4 value of frequency from mobile node 1 to mobile node 2, and from mobile node 2 to mobile node 1 are the same. The relative frequency between two mobile nodes is calculated in such a way that it takes on different emphasis on different segment in the time series based on the *alpha* value. If relative frequency between mobile node *A* and mobile node *B* is equal to 0.4, this means that within the value of 0.4, 20% of the value is supported from the average frequency counted in segment 1, 30% of the value is supported from the average frequency counted in segment 2, and 50% of the value is supported from the average frequency counted in segment 3, since $\alpha_1$=0.2, $\alpha_2$=0.3, $\alpha_3$=0.5.

The concept of relative frequency is best explained by means of an example. In this example, the time series ranges from *t*1 to *t*10, which represents the raw mobile data collected between mobile node *A* and mobile node *B*. The sample of communication collected between two mobile nodes in a mobile environment consists of a sequential order of timeslots, from *t*1 to $t_n$. Each timeslot represents a binary value of either 0 or 1. By having the pre-specified criteria of 3 segments, where segment1 ranges from *t*1 to *t*3 with *alpha* value 0.2, segment2 ranges from t4 to t6 with *alpha* value 0.3 and segment3 ranges from *t*7 to *t*10 with *alpha* value 0.5. The relative frequency between mobile node *A* and mobile node *B* can be presented as an equal:

$$\text{Relative Frequency} = \text{Average(Segment 1)} * \alpha_1 + \ldots + \text{Average(Segment n)} * \alpha_n$$

For the sample data above, a result of relative frequency of 0.93 is obtained. In reality, a frequency of 0.93 would be considered highly logically proximate. The goal of data mining, which is to find relevant and accurate knowledge from raw data, when applied into the mobile environment, would require the time series to be extended to a reasonable amount of length.

After all the relative frequencies between mobile nodes are counted, a sample output would be represented a tabular form represented by Table 2 below. Note that there is no relative frequency within the node itself, and relative frequency is non-directional, thus relative frequency between node *A* and node *B* is equal to relative frequency between node *B* and node *A*.

**Table 1.** Sample Output

| Relative Frequency | Node A | Node B | Node C | Node D |
|---|---|---|---|---|
| **Node A** | - | 0.9 | 0.2 | 0.5 |
| **Node B** | 0.9 | - | 0.5 | 0.8 |
| **Node C** | 0.2 | 0.5 | - | 0.7 |
| **Node D** | 0.5 | 0.8 | 0.7 | - |

Finally, in order to show the knowledge from the raw data, a *frequency threshold* (β) has to be set. The frequency threshold represents the degree of relative frequency that a decision maker is willing to accept in order to recognise two mobile nodes to be logically close to each other. Therefore, within all the relative frequency calculated those which are lesser than β will be treated as not logically close. Therefore, the output is a table or graph where relative frequency ≥ β. From the sample output above, with β = 0.5, there are 7 recognised pair of mobile node.

**Table 2.** Sample Output

| Relative Frequency | Node A | Node B | Node C | Node D |
|---|---|---|---|---|
| **Node A** | - | 0.9 | | 0.5 |
| **Node B** | **0.9** | - | 0.5 | 0.8 |
| **Node C** | | **0.5** | - | 0.7 |
| **Node D** | **0.5** | **0.8** | **0.7** | - |

Graphically, the knowledge is represented in Figure 1. In Figure 1, there are two set of output generated. The left hand side graph represented the output with a *beta* value of 0.5 while the right hand side graph represents the output with a *beta* value of 0.6. The purpose of representing the output in two different graphs is to show the result of increasing the *beta* value.

Here, the knowledge found is that mobile node A is significantly (0.9) related to mobile node *B*. There is also a strong relationship between mobile node B and mobile node *D* (0.8). Finally, there is also a good relationship between mobile node D and mobile node *C* (0.7). We can observe that there is no significant relationship between: node *A* and node *C*, node *A* and node *D*, node B and node C. In different circumstances, the inverse of the output can be used when there is significant relationship between most mobile nodes while the decision maker is interested in knowing finding mobile nodes that are not logically close thus saving resources.

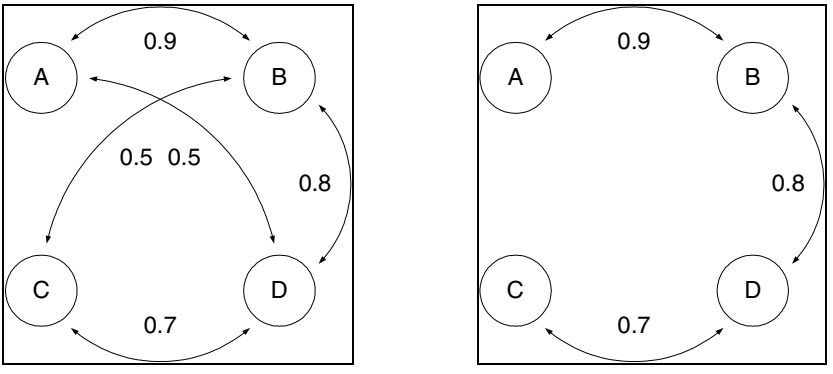


**Fig. 1.** Output with *Beta* = 0.5 (Left), Output with *Beta* = 0.6 (Right)

The algorithm for frequency pattern generation is presented in Figure 2. The frequency generator computes the relative frequency between two mobile nodes with the pre-specified criteria into consideration. The display knowledge part checks and displays only mobile nodes having relative frequency greater than threshold.

```
Function Frequency_Generator (Node A, Node B) {
    Data Structure Representing Pre-Specified Criteria
    Define Configuration As Array Of {
        Array of Segment As Integer
        Array of Segment Size As Integer
        Array of Segment Alpha Value As Float
    }
    Define Freq As Float, I As Integer
    Set Freq To 0 // Each Relative Frequency = Average of Frequency Count * Alpha Value
    For I = 1 To I = Number of Segments
        Freq = Freq + [Average(Segment I) * Alpha(Segment I)]
}
Function Display_Knowledge(Table of Nodes) {
    For I = 1 to I = (No of Mobile Nodes)
        For J = 1 to J = (No of Mobile Nodes)
            If Frequency > Threshold Then Display Mobile Node I-J
}
```

**Fig. 2.** Frequency Generator Algorithm

## 4 Performance Evaluation

Performance evaluation is performed by measuring how accurate the proposed solution can determine three different sets of data which contains different kind of frequency distribution. The three sets of data consists of Set *A*, Set *B* and Set *C*. Set *A* consists of a distribution of frequency skewed to the left. Set *B* consists of a near

normal distribution of frequency. Set *C* consists of a skewed right distribution of frequency. The sample data sets are first generated, and the formula entered to calculate the relative frequency based on the pre-specified criteria which consist of segments and alpha values.

Figure 5 below represents the impact of increasing emphasis on recent communication to the relative frequency. The figure shows that as there is increasing emphasis on recent communication, Set *C*, the set of data which has more recent communication tends to have higher relative frequency, which will be regarded as having more significant logical relationship.

Figure 3 represents the performance result of increasing emphasis on recent and historical data. The higher the window, the longer the decimal places required. From the performance data, we can observe clearly that Set C responded strongly to recent emphasis and Set A responded strongly to emphasis on historical data.

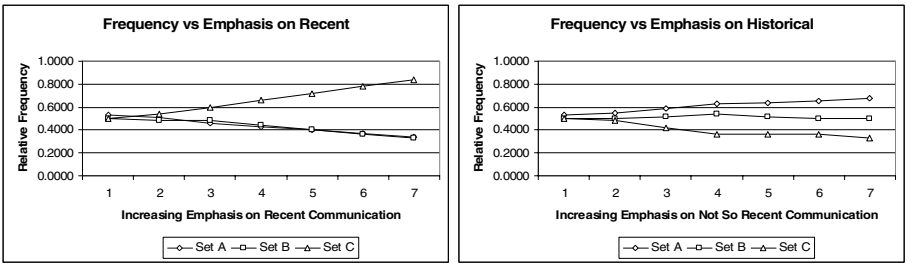


**Fig. 3.** Performance Data for Recent & Historical Emphasis

Figure 4 below represents the impact of decreasing standard deviation of emphasis on relative frequency. The decreasing standard deviation also means that more emphasis will be placed on the middle region of the pre-specified criteria. As Set *B* contains a normally distribution frequency distribution, it stays within the average region. Set *A*, however only achieved a slight increase while Set *C* decreases significantly.

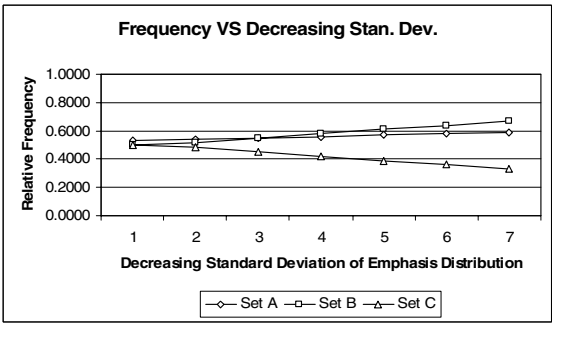


**Fig. 4.** Performance Data for Decreasing Standard Deviation

## 5 Conclusion and Future Work

The proposed frequency pattern [4] is more relevant in mining knowledge out of mobile users in real life. This is because the frequency pattern focuses on the logical communication as a mean to determine whether two mobile nodes are in fact close to each other logically rather than using physical distance which may be more appropriate in a non-mobile environment. The frequency pattern uses the concept of segments and *alpha* values in order to allow decision makers to place different level of emphasis during different parts of the time series, such as during peak hour, or during hours that should be neglected such as out of office hour.

The performance of the proposed frequency pattern [4] method significantly reacts to the data set. It can further be tailored into different environment by dynamically changing the size of the segment, the alpha value and the number of segments. The future work of this proposed paper is to looking at the relative relationship between mobile nodes in one further step. It may be possible to find out the relationship between two mobile nodes through an intermediate node.

## References

1. R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules. In *Proc. 20th Int. Conf. Very Large Data Bases*, 1994.
2. E.-P. Lim, et al. In Search Of Knowledge About Mobile Users. 2003.
3. R. T. Ng and J. Han. Efficient and Effective Clustering Methods for Spatial Data Mining. In *20th International Conference on Very Large Data Bases, September 12--15, 1994, Santiago, Chile proceedings*, 1994.
4. Y. Wang, E.-P. Lim, and S.-Y. Hwang. On Mining Group Patterns of Mobile Users. *Lecture Notes in Computer Science*, vol. 2736, pp. 287-296, 2003.