Performance Evaluation of Linkage Pattern Mining Method

Saerom Lee, Yusuke Okubo and Yoshifumi Okada
Division of Production and Information Systems Engineering, College of Information and Systems
Muroran Institute of Technology, 27-1, Mizumoto-cho, Muroran, Hokkaido 050-8585, Japan
saerom@cbrl.csse.muroran-it.ac.jp, ookubo@cbrl.csse.muroran-it.ac.jp and okada@csse.muroran-it.ac.jp

ABSTRACT

Linkage pattern mining is a data mining technique for discovering sets of patterns that appear repeatedly across multiple sequential data. We previously proposed a linkage pattern mining method using closed itemset mining and showed that it can effectively exclude pseudo patterns derived from noise. In addition, it was suggested that the extraction accuracy of this method was strongly affected by the correctness of frequent pattern extraction from each sequential data. The frequent pattern extraction process requires two parameters, a window width \( w \) and a minimum number of occurrences \( \theta \). However, our previous study has not performed comprehensive evaluation in various combinations of these parameters. In this study, we conducted a grid search for the parameter values that exhibit stable and high extraction accuracy. As a result, it is shown that both parameters should be set to smaller values.

KEYWORDS

linkage pattern mining, sequential pattern mining, parameter, grid search

1 INTRODUCTION

Sequential pattern mining is a promising and effective data mining method for finding frequent patterns in large-scale sequential data. To date, many algorithms have been developed [1]–[3] and applied in various fields such as E-commerce [4]–[7] and life science [8]. Research into sequential pattern mining can be broadly classified into two types: approaches that target single sequential data and those that target multiple sequential data. The former aims to find repeating and frequently occurring patterns (frequent patterns or episodes) in sequential data [9]–[13]. The latter focuses on detecting same or similar subsequences among sequential data [14]–[16].

We previously have proposed linkage pattern mining method [17, 18] as one of the sequential pattern mining techniques. Fig. 1 shows an example of a linkage pattern mining. As shown in this figure, the linkage pattern mining method finds a set of patterns that repeat across multiple sequential data. Even if patterns do not show similarity among sequential data, the set of those patterns is extracted as a linkage pattern if those patterns occur continually within the same period.

In recent our study [18], we introduced a closed itemset mining to the original method [17] to enhance the extraction accuracy by excluding pseudo patterns derived from noise.
This idea allowed more noise-robust linkage pattern mining than the original method. On the other hand, it was suggested that the extraction accuracy of this method was strongly affected by the correctness of frequent pattern extraction from each sequential data. The frequent pattern extraction process requires two parameters, a window width $w$ and a minimum number of occurrences $\theta$. However, our previous study [18] has not performed comprehensive evaluation in various combinations of these parameters.

The aim of this study is to identify adequate parameter values that exhibit stable and high extraction accuracy. In this paper, we show results of a grid search for these parameter values. The remainder of this paper is organized as follows. Section II explains to procedure of the linkage pattern mining method. Section III explains the method of the grid search for the parameter values. Section IV presents the experimental results and discusses some observations. Section V provides an overall summary.

2 DEFINITION OF LINKAGE PATTERN

Suppose that multiple sequential data are given as input, and frequent patterns have already been extracted from those sequential data. A linkage pattern is defined as the set of frequent patterns that satisfy the following conditions in an interval over those sequential data:

1) For all the frequent patterns appearing in the interval, there exist one or more frequent patterns overlapping for the occurrence intervals.  
2) A set of the frequent patterns that satisfy the condition 1) occurs $\alpha$ or more times along the sequential data.

3 METHOD

Fig. 2 shows the procedure of the proposed method. Fig. 2a, 2b, and 2d are the steps implemented in the previous method: extracting and labeling frequent patterns from each sequence (Fig. 2a), generating interval graphs depending on overlapping labels on the time axis (Fig. 2b), and outputting the linkage pattern (Fig. 2d). In this method, a new step (Fig. 2c) is introduced, i.e., closed itemset
mining from the generated interval graphs. This resolves the problem by which linkage patterns are contaminated by noise data, as observed in the previous method. These steps are explained in detail below.

3.1 Frequent pattern extraction and labeling

First, normalization and discretization are executed on all sequential data in a preprocess. In the normalization, sequential data are converted to a scale from 0 to 1. In the discretization, the range of normalized data (0–1) is divided at the \( D \) stages, and a discrete value from 0 to \( D-1 \) is allocated to each data.

Next, repeatedly occurring frequent patterns are extracted from the sequential data using Mannila’s algorithm [13]. This algorithm uses a window width \( w \) and minimum number of occurrences \( \theta \) as input parameters, where \( w \) and \( \theta \) are natural numbers \( \geq 2 \). Window width \( w \) is the length of the slice used to scan sequential data. Frequent patterns are searched by increasing the window width in increments of 1 in range from \( \text{ceil}(w/2) \) to \( w \), where \( \text{ceil}(x) \) is the ceiling function that returns the smallest integer not less than \( x \). Namely, frequent pattern length is equal to or greater than \( \text{ceil}(w/2) \).

The minimum number of occurrences \( \theta \) is the minimum number of frequent patterns to be extracted. Mannila’s algorithm finds frequent patterns that satisfy \( \theta \) for a specified \( w \).

The labeling process applies the same label to the same frequent pattern. This process is performed after excluding frequent patterns with length less than \( w/2 \). When multiple frequent patterns occur within the same sequential data and same periods, labeling is performed for the maximum length frequent pattern.

3.2 Interval graph generation

Here, a labeled frequent pattern is referred to as a label. In this step, interval graphs are generated from the interval representation of each label. An interval graph is obtained by associating each label with a node and an overlap of any two labels on the time axis between sequential data with an edge [19]–[21]. In other words, an interval graph is a set of frequent patterns that occur in a linked manner in the same period between different sequential data.

The previous method outputs the interval graph with the highest frequency as a linkage pattern. However, frequent patterns that are accidentally constructed by noise (pseudo patterns) cause the following problems. If different pseudo pattern labels are attached to the same interval graphs, these interval graphs are considered completely different despite having an identical linkage pattern. This reduces the accuracy of linkage pattern mining.

3.3 Extraction of linkage patterns based on closed itemset

Pseudo patterns tend to occur randomly on the time axis; thus, the probability that the same pseudo pattern will be included in multiple equivalent interval graphs is extremely low. Therefore, it is expected that pseudo patterns can be excluded by extracting label sets that occur commonly in multiple interval graphs. In our method, pseudo patterns are excluded by mining closed itemsets on the obtained interval graphs.

Definition of closed itemset is shown below. Let \( I = \{1, 2, \ldots, n\} \) be a set of items. A transaction database on \( I \) is a set \( T = \{t_1, t_2, \ldots, t_m\} \) such that each \( t_i \) is included in \( I \). Each \( t_i \) is called a transaction. A set \( P \subseteq I \) is called an itemset. A transaction including \( P \) is called an occurrence of \( P \). The set of occurrences of \( P \) is expressed as \( T(P) \). The size of a set of occurrences for \( P \) is referred to as the frequency of \( P \). An itemset \( P \) is called a closed itemset if no other itemset \( Q \) satisfies \( T(P) = T(Q) \), \( P \subseteq Q \). For a given minimum support constant
(hereafter \( \text{minsup} \)), \( P \) is frequent if \( |T(P)| \geq \text{minsup} \). A frequent and closed itemset is referred to as a frequent closed itemset.

Fig. 2c shows the process of excluding pseudo patterns from interval graphs. Each interval graph is considered a transaction, and each node in the interval graph is considered an item. By applying closed itemset mining to this transaction database, we can extract the maximal node sets (closed itemsets) that are shared in \( \text{minsup} \) or more interval graphs. Finally, the closed itemset with the highest frequency is output as the linkage pattern. Thus, it is possible to extract linkage patterns with greater accuracy as randomly constructed pseudo patterns can be excluded. Fig. 2c illustrates an example of how pseudo patterns \( n_A, n_B, \) and \( n_C \) are excluded; only the authentic linkage patterns \( \{A, B, C\} \) are extracted. In this study, we use the fast and exhaustive linear closed itemset miner (LCM) algorithm [22].

4 EXPERIMENTS

In this section, the manner of the grid search for the two parameters, \( w \) and \( \theta \), is explained. The grid search is performed using artificially created five sequential datasets. The aim of this experiment is to find good parameter values leading to high extraction accuracy.

4.1 Artificial datasets

To find good parameters, it is necessary to test various datasets. In this study, we conducted experiments using five artificially created sequential datasets in which synthetic linkage patterns were embedded in different manners.

Each artificial dataset was composed of three sequential data. The sequential data were generated by inserting 10 linkage patterns (embedded linkage patterns) into random sequential data created using uniform random numbers. For this experiment, we created five artificial datasets (Dataset1–Dataset5) that included noise in embedded linkage patterns were created by adding fluctuations to each
Figure 4. Extraction accuracies in different combinations of \( w \) and \( \theta \)
time point in the linkage patterns. The fluctuations were generated using normal random numbers (SD = 0.01). Fig. 3 shows a section of each artificial dataset. The formats of linkage patterns embedded in each dataset are as follows. Dataset1 is an artificial dataset wherein equal length frequent patterns were embedded with the same start time across the three sequential data (Fig. 3a). Dataset2 is an artificial dataset wherein equal length frequent patterns were embedded with different start times across the three sequential data (Fig. 3b). Dataset3 is an artificial dataset wherein different length frequent patterns for each of the three sequential data were embedded at the same time (Fig. 3c). Dataset4 is an artificial dataset wherein frequent patterns with different lengths for each of the three sequential data were embedded at different times (Fig. 3d). Dataset5 is an artificial dataset wherein one or two types of frequent patterns were embedded with different lengths and different start times for each of the three sequential data (Fig. 3e).

4.2 Parameter settings

Minimum numbers of occurrences ($\theta$) were set to 5, 10, 15, 20, 25 and 30, and window widths ($w$) were set to natural numbers in the range $3 \leq w \leq 10$. The range of $w$ was determined based on the following. When $w = 2$, each data point will be labeled; thus, meaningless frequent patterns that consist of only one data point are generated. In case of $w > 10$, the results are not shown because it requires much computation time in each combination of parameters. The minsup in closed itemset mining was set to 5.

4.3 Extraction accuracy of linkage patterns

The extraction accuracies of the embedded linkage patterns for the previous and proposed methods were compared using the above 5 artificial datasets. Precision, recall, and F-measure were used as evaluation indexes. These indexes were calculated as follows.

$$ \text{Precision} = \frac{\text{CDP}}{\text{DDP}} $$

$$ \text{Recall} = \frac{\text{CDP}}{\text{EDP}} $$

$$ \text{F-measure} = 2 \cdot \text{Precision} \cdot \text{Recall} / (\text{Precision} + \text{Recall}) $$

Here, CDP is the number of data points in the correctly detected areas of the embedded linkage patterns, DDP is the number of data points in the areas of the embedded linkage patterns detected by the method, and EDP is the number of data points in the embedded linkage patterns.

5 RESULTS AND DISCUSSION

Fig. 4 is a result of the grid search for the two parameters, $w$ and $\theta$. These graphs show precision, recall, and F-measure for each dataset. We can see that extraction accuracy indicates a similar tendency among all the datasets.

Precision shows scores $\geq 80\%$ in most of ($w$, $\theta$). However, when two parameters take large $w$ and small $\theta$, the scores are low. This reason is as follows. By larger $w$, more frequent patterns are extracted, because frequent patterns of lengths in the range from 2 to $w$ are exhaustively searched. This means that it increases the chance of extracting false linkage patterns. On the other hand, smaller $\theta$ also lead to increase of false linkage patterns, because many frequent patterns with low frequencies are extracted.

Recall tends to decrease with increasing $w$ in all the datasets. In particular, rapid decreases of the scores are observed in $w \geq 5$. This is because the number of frequent patterns extracted from each sequence decreases dramatically in $w \geq 5$. Therefore, the number of obtained interval graph is also reduced dramatically. On the other hand, the scores of recall show gradual decrease with increasing $\theta$. In addition, drastic fluctuations of the scores by $\theta$ are not observed in all the datasets. This means that recall is affected mainly by value of $w$ than $\theta$. 

ISBN: 978-1-941968-17-8 ©2015 SDIWC
From the above results, we can see that extraction accuracy is strongly affected by value of \( w \). Our method extracts frequent patterns whose lengths are equal to or greater than \( \text{ceil}(2/w) \) as described in section 3.1. Use of too large \( w \) obstructs extracting linkage patterns composed of short frequent patterns. In contrast, small \( w \) enables us to extract such linkage patterns. In addition, \( w \) and \( \theta \) should be set to smaller values to obtain higher extraction accuracy.

### 6 CONCLUSION

Extraction accuracy of our linkage pattern mining is strongly affected by the correctness of the frequent pattern extraction. The aim of this study was to find parameter values of the frequent pattern extraction process so that our linkage pattern mining method exhibits stable and high extraction accuracy. In this paper, we conducted a grid search in different combinations of the two parameters (window width \( w \) and minimum number of occurrences \( \theta \)) in the frequent pattern extraction process using five artificial datasets. In this grid search, the extraction accuracies were evaluated using three indexes, precision, recall and F-measure. As a result, it was shown that use of smaller \( w \) and \( \theta \) enables higher extraction accuracy.

In future, we will investigate the impact of the parameter \( \text{minsup} \) in the step of closed itemset mining. Furthermore, we will evaluate using more various artificial datasets to ensure our suggestions regarding the adequate parameters. In addition, we will address increasing the speed of the frequent pattern mining algorithm and apply the method to large-scale real sequential data with noise/fluctuations, such as vital data and crustal movement data.

### REFERENCES


