Extraction of Frequent Association Patterns Co-occurring across Multi-sequence Data

Takahiro Miura and Yoshifumi Okada

***Abstract*— The progress in computer performance and ubiquitous sensor technology has made it possible to yield a large amount of data such as vital data and earthquake data. In general, sensor data are measured from multiple observation points. Extracting correlation or causal association from such huge multi-sequences is a challenging issue in the field of data-mining technology. Development of high-performance sequence mining tools will make great contributions for analyzing various real data such as disease risk prediction or earthquake prediction. In this paper, we propose a new method for extracting sets of patterns (called *association patterns*) that co-occur repeatedly across multiple sequences. Extraction of association patterns is performed by a combination of frequent pattern extraction and interval graph mining. Our method is different from traditional multi-sequence mining methods in that it does not assume similarity of patterns among different sequences. Namely, even if frequent patterns in different sequences show no similarity, our method extracts them as an association pattern if these patterns exhibit a frequent co-occurrence relation along a time-sequence. In this paper, we evaluate the usefulness of our method using synthetic datasets.**

***Index Terms*— frequent pattern, association pattern, interval graph, data mining**

I. Introduction

The progress in computer performance and ubiquitous sensor technology has made it possible to yield a large amount of data such as vital data and earthquake data. In general, sensor data (typically time sequence data) are measured from multiple observation points. Extracting correlation or causal association from such huge multi-sequences is a challenging issue in the field of data-mining technology. Development of high-performance sequence mining tools will make great contributions for analyzing various real data such as disease risk prediction or earthquake prediction.

So far, many data-mining approaches have been proposed to discover frequent patterns (or motifs) for a single sequence data [1]-[3]. Other several research groups suggested methods for discovering similar patterns among multi-sequences and applied them to sensor fault detection or server load balancing [4], [5].

Manuscript received December 30, 2011. This work was supported in part by JUTEN KENKYU PROJECT form Muroran Institute of Technology.

T. Miura is with the Department of Information and Electronic Engineering, Muroran Institute of Technology, 27-1, Mizumoto-cho, Muroran 050-8585, Japan (e-mail: miura@cbrl.csse.muroran-it.ac.jp).

Y. Okada is with College of Information and Systems, Muroran Institute of Technology, 27-1, Mizumoto-cho, Muroran 050-8585, Japan(corresponding author to provide phone: +81-143-5408; fax: +81-143-5408; e-mail: okada@csse.muroran-it.ac.jp)

In this paper, we propose a new method for extracting sets of patterns (called *association patterns*) that co-occur repeatedly across multiple sequences. Extraction of association patterns is performed by a combination of frequent pattern extraction and interval graph mining. Our method is different from traditional multi-sequence mining methods in that it does not assume similarity of patterns among different sequences. Namely, even if frequent patterns in different sequences show no similarity, our method extracts them as an association pattern if these patterns exhibit a frequent co-occurrence relation along a time-sequence. In this paper, we evaluate the usefulness of our method using synthetic datasets.

This paper is organized as follows. Section II describes the basic idea of the method. Section III explains the procedure of our method. Section IV and V show the experimental method and the results, respectively. Finally, Section VI summarizes our conclusions and suggests future work.

II. The Basic Idea of Our Method

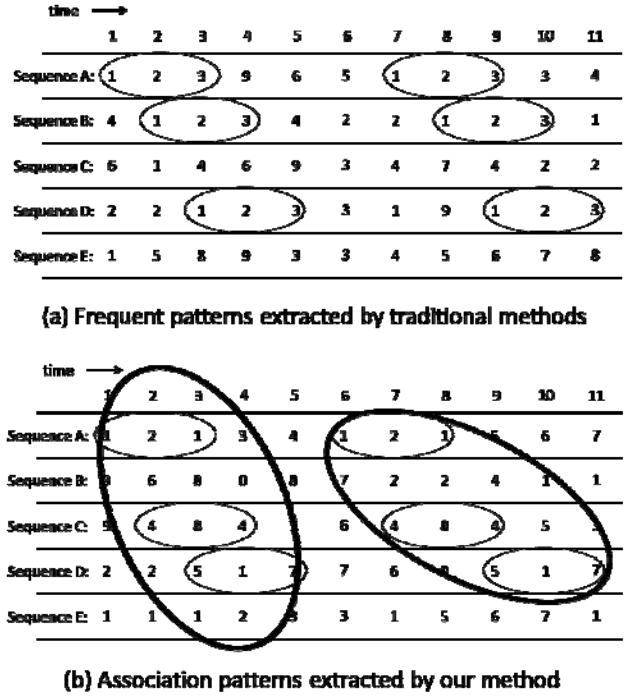
Fig. 1(a) summarizes traditional pattern mining for a single sequence [1]-[3] and for multiple sequences [4], [5]. Methods for a single sequence extract patterns that occur repeatedly in a sequence. On the other hand, methods for multiple sequences discover patterns that are similar or common among different sequences.

In contrast to those traditional approaches, we aim at discovering sets of patterns that co-occur repeatedly across multiple sequences, as shown in Fig.1b. Respective patterns in a set do not necessarily need to be identical. In this paper, such set of patterns is referred to as an *association pattern*.
We first apply traditional frequent pattern mining to each sequence and subsequently employ a concept of interval graph [3] to extract association pattern. An interval graph is a graph for representing a set of intervals as depicted in Fig. 2(a) and Fig. 2(b) is the interval graph for Fig. 2(a), in which a node indicates an interval and an edge means that two intervals overlap. In this study, a frequent pattern in each sequence is regarded as a node, and an overlap of any two frequent patterns is considered as an edge. Extraction of association patterns can be achieved by finding connected graphs from the interval graph generated above.

III. Method

Fig. 3 illustrates the procedure of the method. This method consists of two steps: (a) frequent pattern extraction from each sequence and (b) association pattern extraction from the interval graph. These steps are performed after preprocessing for each sequence.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)



(a) Frequent patterns extracted by traditional methods

(b) Association patterns extracted by our method

Fig. 1 : Difference between traditional methods and our method

A. Preprocessing for each sequence

Each sequence is normalized and discretized as follows. First, data x_t in a time t of a sequence is normalized into a range from 0 to 1 by the following equation:

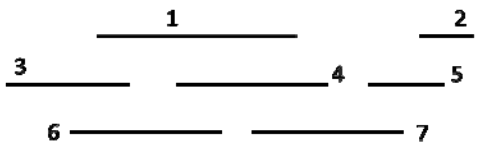
$$normalize(x_t) = \frac{x_t - \min}{\max - \min},$$

where *max* and *min* are the maximum value and the minimum value in the sequence, respectively.

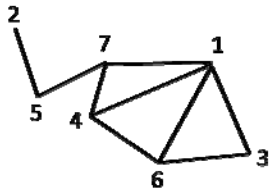
Subsequently, every normalized data is represented by a discretized value, dividing uniformly the range from 0 to 1 into D-grades.

B. Frequent pattern extraction from each sequence

Fig. 3(a) exhibits frequent pattern extraction from each sequence and labeling process for those. First, frequent patterns are extracted from each sequence. In this study, we employed a frequent pattern mining algorithm proposed by Mannila *et al.* This algorithm requires two input parameters: a maximum window width w of patterns extracted and a frequency threshold θ. Next, each mined frequent pattern is tagged with a unique label according to the kind of pattern. In this process, frequent patterns whose lengths are 1 are eliminated without labeling. In addition, if there are multiple frequent patterns within a window, a maximal frequent pattern among them is labeled. Finally, an interval representation (*i.e.*, interval graph) of frequent patterns derived from every sequence is generated.



(a) Interval representation



(b) Graph representation

Fig. 2 : Interval graph

C. Association pattern extraction from the interval graph

Fig. 3(b) exhibits association pattern extraction from interval graph. In this step, sets of intervals overlapping among sequences are searched to obtain association patterns. Extraction of association patterns can be done in $O(n)$ time because we just seek for labeled frequent patterns along n time points. Finally, each association pattern is output as a set of labeled frequent patterns.

IV. Experiments

In this study, we evaluate the extraction accuracy of association patterns using synthetically generated dataset. A dataset is composed of three synthetic sequence data, and each sequence consists of ten same patterns that are artificially embedded to random values derived from a uniform distribution. In this experiment, we create three types of datasets: Dataset1, Dataset2 and Dataset3. Dataset1 is a dataset that includes an identical frequent pattern in same timings in the three sequences. Dataset2 is a dataset in which an identical frequent pattern is embedded in different timings among the three sequences. Dataset3 is a dataset in which frequent patterns with different length (5, 11 and 15 points) appear in same timings in the three sequences. Figure4 shows the sequence data for Dataset3.

In this experiment, the maximum window width w is set to 5, 10 and 20, and the frequent threshold is set to $\theta = 10$. The performance of the method is evaluated using extraction accuracy of embedded association patterns. We employed Precision and Recall as evaluation indices. These indices are calculated by the following equations:

$$Precision = CDP / DDP,$$
$$Recall = CDP / EDP.$$

CDP is the number of data points that are correctly detected from embedded frequent patterns. *DDP* is the number of data points that are obtained by our method. *EDP* is the number of data points of embedded frequent patterns.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

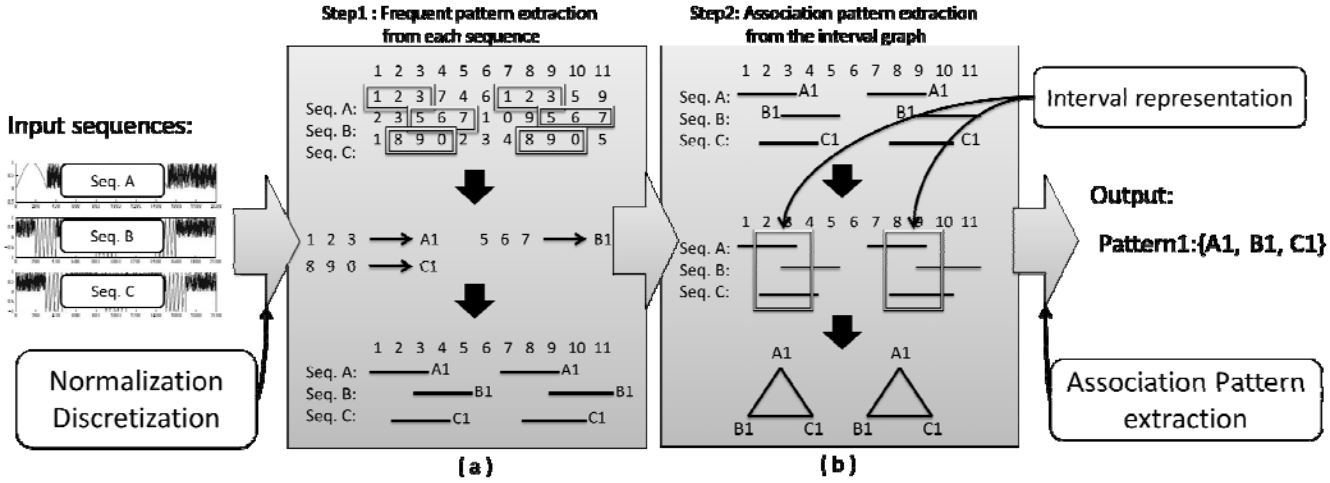


Fig. 3 : The procedure of our method

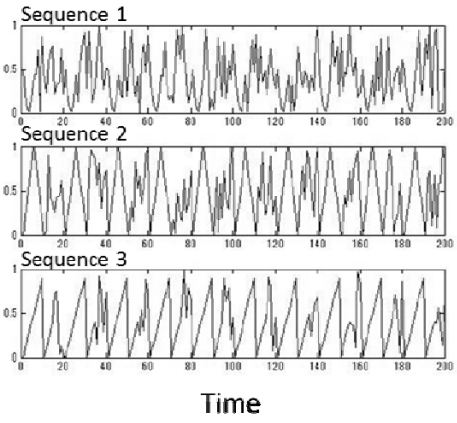
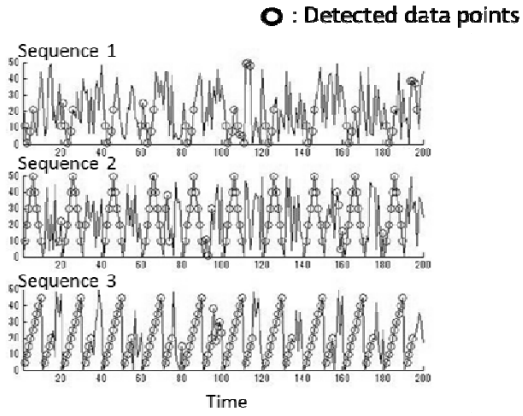


Fig. 4 : A example of synthetic data (Dataset3)



Examples of association patterns
(A, B and C correspond to Sequence1, 2 and 3, respectively)
Num 1 : {A10, A6, B18, B8, B19, B9, C35, C18, C19}
Num 2 : {A10, A6, B19, B9, B7, C19, C35}
Num 3 : {A10, A6, B18, B8, B19, B9, B6, C35, C18, C19}
Num 4 : {A10, A6, B18, B8, B19, B9, C20, C35, C18, C19}
Num 5 : {A10, A6, A9, A7, B18, B8, B19, B9, C35, C18, C19}
Num 6 : {B15}

Fig. 5 : A example of association pattern extraction

V. RESULTS AND DISCUSSION

As an example, we show the detection result of association patterns (open dots) on Dataset3 in Fig. 5. The set representations of the detected association patterns are shown in the bottom of the figure in which the correct detection and false detection are denoted in bold and plain text, respectively. From this figure, we see that the embedded association patterns are adequately extracted, whereas non-embedded association patterns are also detected falsely. Most of the errors are caused by pseudo patterns created by random values, and fragmentation of embedded association patterns.

Fig. 6 and Fig. 7 are graphs of the *precision* and *recall*, respectively. Each graph presents the scores of respective window widths (5, 10 and 20). We can see that the *precisions* decrease with an extension in window width in all the datasets. This is because larger window captures the area of non-embedded patterns. In contrast, the recalls exhibit an upward trend with an extension in the window width. It means that larger window width can widely cover the embedded patterns. In general, there is a tradeoff between *precision* and *recall*. It will become important to balance these two indices by estimating the distribution of lengths of candidate frequent patterns.

VI. CONCLUSION

We proposed a method for discovering association patterns that co-occur repeatedly across multiple sequence data. This method extracted association patterns by detecting connected graphs from an interval graph for the frequent patterns appearing in each sequence. We applied it to three synthetic datasets and evaluated its performance using *precision* and *recall*. As a result, embedded association patterns were extracted with a high degree of accuracy. Furthermore we found that employing an appropriate window width in each dataset is important to enhance the performance of the method.

In the future, we will develop a method for estimating proper parameters according to the distribution of lengths of candidate frequent patterns. In addition, we will evaluate the method using not only more noisy datasets with stochastic fluctuations but also real datasets such as vital data and earthquake data.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

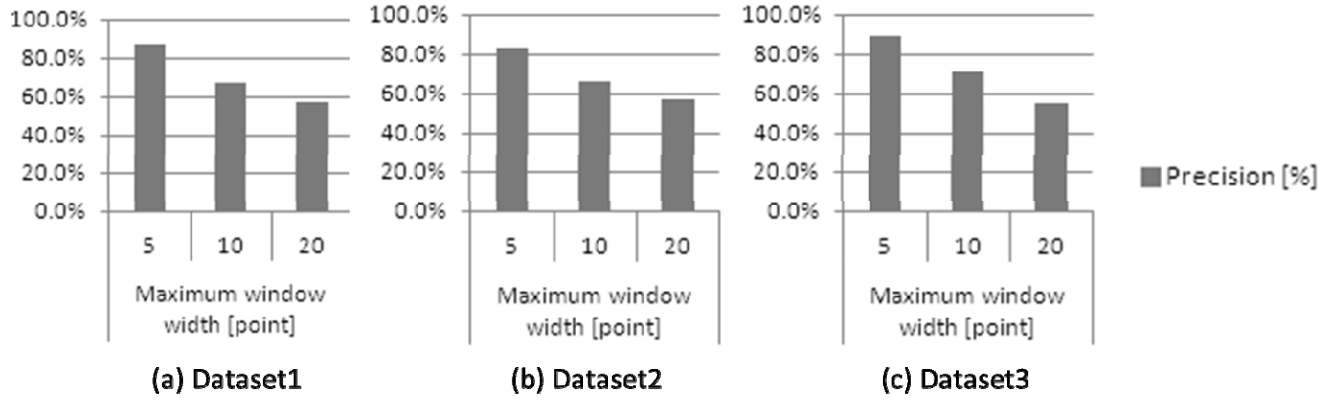


Fig. 6 : Precision

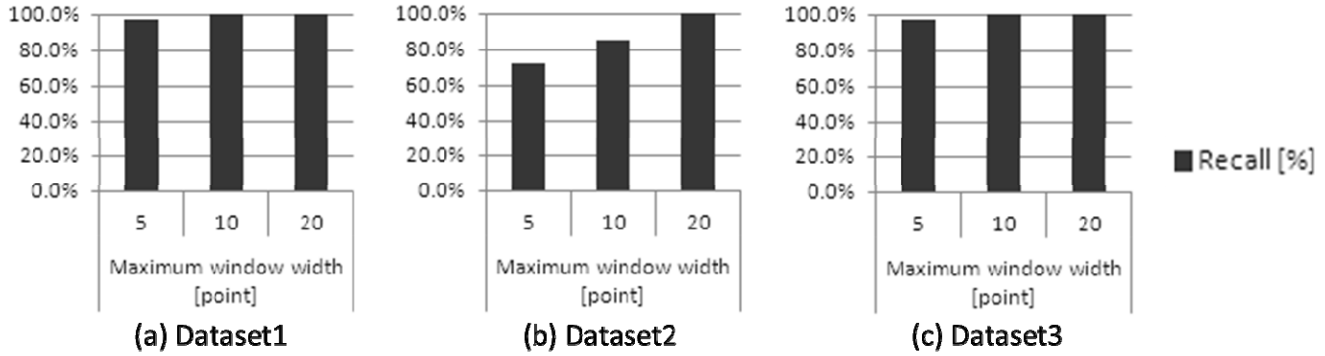


Fig. 7 : Recall

ACKNOWLEDGMENT

This work was supported in part by JUTEN KENKYU PROJECT 2011 form Muroran Institute of Technology.

REFERENCES

[1] Q. Zhao and S. S. Bhowmick, “Sequential Pattern Mining: A Survey”, Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003118, 2003.
[2] R. Agrawal and R. Srikant, "Mining Sequential Patterns", Proc. of The 11th Int'l Conf. on Data Engineering , pp.3-14, 1995.
[3] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu, "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth", Proc. of The 17th Int'l Conf. on Data Engineering, pp.215-224, 2001.
[4] Y. Zhu and D. Shasha, “StatStream: Statistical Monitoring of Thousands of Data Streams in Real Time”, Proc. of VLDB, pp.358-369, 2002.
[5] Y. Sakurai, S.Papadimitriou and C. Faloutsos, “BRAID: Stream Mining through Group Lag Correlations”, Proc. of ACM SIGMOD Conference, pp.599-610, 2005.
[6] N. Korte and R. H. Mohring, “An incremental linear-time algorithm for recognizing interval graphs.”, SIAM Journal on Computing, vol. 18, pp.68-81, 1989.
[7] G. S. Lueker and K. S. Booth, “A linear time algorithm for deciding interval graph isomorphism.”, Journal of the ACM, vol. 26, pp.183-195, 1979.
[8] H. Mannila, H. Toivonen and A. I. Verkamo, “Discovery of Frequent Episodes in Event Sequences.”, Data Mining and Knowledge Discovery 1, pp.259-289, 1997.