Linkage Pattern Mining based on Causal Relationship

Yusuke Okubo, Saerom Lee, and Yoshifumi Okada

***Abstract*—We have previously developed a method for extracting linkage patterns, a set of patterns that appear repeatedly across multiple sequential data. In this study, we propose a new linkage pattern mining method based on causal relationships among such patterns. We assume that the user is interested in events appearing in particular sequential data (target sequence) from multiple sequential data. The proposed method extracts linkage patterns showing causal relationship by identifying patterns inducing the events of interest from the other sequences (non-target sequences). The proposed method can improve the problems in the previous method, which extracts pseudo linkage patterns. The proposed method was applied to artificial sequential datasets, and extraction accuracy was compared with the previous method. Experimental results show that the proposed method can extract linkage patterns showing causal relationships with high accuracy compared with that of the previous method.**

***Index Terms*—causal relationship, linkage pattern, sequential pattern mining**

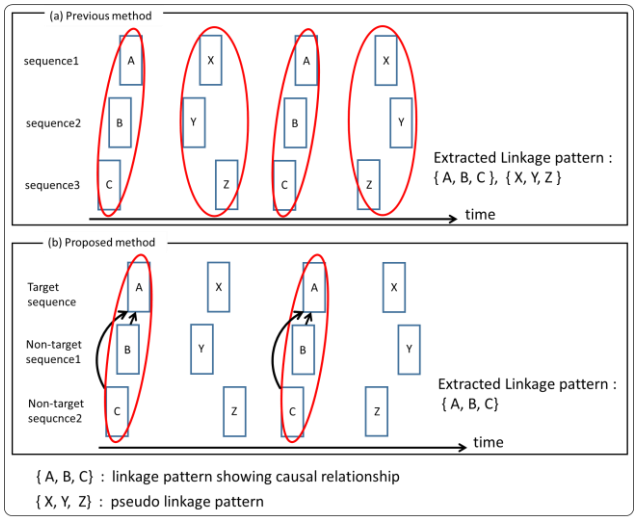

Fig. 1. Previous method and proposed method

I. Introduction

In recent years, sequential pattern mining has attracted attention as a powerful technique to discover useful information and knowledge from a large amount of sequential data [1]–[8]. We have previously developed an original sequential pattern mining method to extract linkage patterns from multiple sequential data. Figure 1(a) shows an example of linkage pattern mining. A linkage pattern is defined as a set of patterns that occur consecutively within the same period across multiple sequential data and appear repeatedly along the sequential data.

In this study, we propose a new linkage pattern mining approach based on causal relationships among patterns. Figure 1(b) shows an example of the proposed linkage pattern mining. We assume that user is interested in events (patterns) appearing in particular sequential data from multiple data.

Here, the particular sequential data is referred to as a target sequence, and the appearing events are referred to as target patterns. The other sequential data are called non-target sequences, and the events appearing in them are referred to as non-target patterns. Our goal is to extract linkage patterns that show causal relationship, wherein a target pattern is induced by non-target patterns. In an extracted linkage pattern, the target pattern and non-target patterns are referred to as a result pattern and cause patterns, respectively. A causal relationship among patterns is represented by a directed weighted graph. Using this method, we expect that the cause of interesting events can be discovered from multiple sequential data such as vital data [9]–[13] or crustal movement data [14]–[15].

In this study, we apply the proposed method to artificial datasets and compare extraction accuracy to that of the previous method.

Manuscript received January 5, 2016.

Y. Okubo is with the Division of Information and Electronic Engineering, Muroran Institute of Technology, 27-1, Mizumoto-cho, Muroran, Hokkaido 050-8585, Japan (e-mail: ookubo@cbrl.csse.muroran-it.ac.jp). 2015.

S. Lee is with the Division of Production and Information Systems Engineering, Muroran Institute of Technology, 27-1, Mizumoto-cho, Muroran, Hokkaido 050-8585, Japan (e-mail: saerom@cbrl.csse.muroran-it.ac.jp).

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

II. Method

Figure 2 shows the procedure of the proposed method.

A. Input dataset

The input to the method is a set of sequences, as shown in Fig. 2. In addition, for each sequence, the user specifies a target or non-targets, where the target is set to one of those sequences, and the non-targets are set to the others. In Fig. 2, four sequences are used as an input, in which the target is represented by T, and the non-targets are denoted N_1, N_2, and N_3. Each frequent pattern appearing in the target is called a target pattern and is denoted t_k, where k is an index of the target pattern. Each frequent pattern appearing in a non-target is called a non-target pattern and is denoted n_{lm}, where l is an index of the non-target sequence and m is an index of the non-target pattern.

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

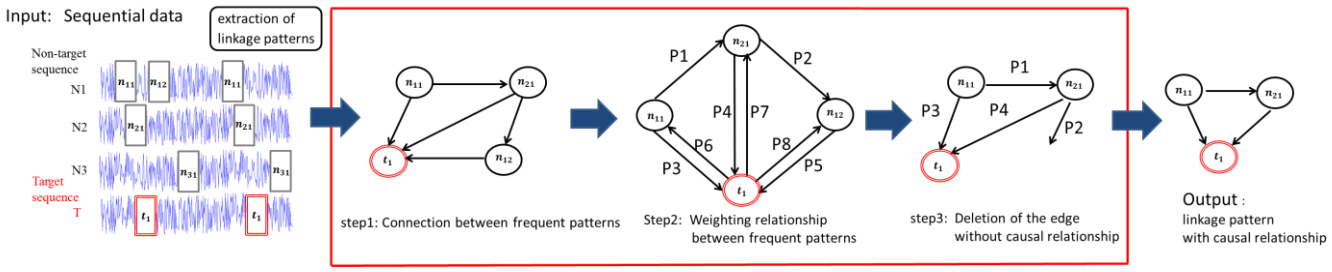

Fig. 2. Procedure of proposed method

B. *Directed graph of frequent patterns*

Linkage patterns are extracted in advance from the input sequences using the previous method [16]. Subsequently, the frequent patterns (target patterns and non-target patterns) within each linkage pattern are connected by directed edges (Step 1, Fig. 2). The connections of edges between the frequent patterns consist of the following.

- Connection between non-target patterns
- Connection between target patterns and non-target patterns

Non-target patterns are connected by directed edges from a prior pattern to a posterior pattern in the order of those occurrences. Each target pattern is connected with every non-target pattern by a bidirectional edge.

C. *Weighed directed graph of frequent patterns*

The edges between frequent patterns are weighted as follows (Step 2, Fig. 2). Here, let *x* and *y* be frequent patterns and assume an edge is connected from *x* to *y*. The probability that the edge is connected from *x* to *y*, p(x→y), is estimated as follows:

$$P(x \rightarrow y) = Occ(x \rightarrow y) / Occ(x). \quad (1)$$

Here, Occ(*x*) is the number of occurrences of *x* and Occ(*x*→*y*) denotes the frequency that the edge is connected from *x* to *y*.

D. *Deletion of the edge without causal relationship*

If the value of formula (1) is less than *thr*, the edge between those patterns is deleted (Step 3, Fig. 2). *thr* is a threshold to discriminate a causal relationship between a target pattern and a non-target pattern. The remaining target patterns and non-target patterns are considered the result patterns and the cause patterns, respectively.

E. *Output of linkage pattern with causal relationship*

This method outputs each linkage pattern that consists of a set of the result/cause patterns and a set of edges that represents a causal relationship.

III. Evaluation Experiment

A. *Artificial dataset*

Our artificial datasets are composed of four sequences, T, N_1, N_2, and N_3, and are created as follows. First, four random sequences are created by generating 4000 uniform random numbers for each sequence. Next, 20 linkage patterns (true linkage patterns) showing causal relationship are embedded across the four random sequences. In a true linkage pattern, one result pattern is placed in T, and three cause patterns are embedded sequentially into N_1, N_2, and N_3. Subsequently, 20 non-target patterns that are not included in any true linkage patterns are embedded randomly into each of N_1, N_2, and N_3.

By the above operations, we created six artificial datasets (test1–test6). Table 1 shows the details of the artificial datasets, and Fig. 3 shows an extraction from test6. Note that 10 dummy linkage patterns that are partially excerpted from the true linkage patterns were embedded randomly into test5 and test6.

B. *Evaluation of extraction accuracy*

We applied the proposed method to the six datasets to investigate extraction accuracy considering the embedded true patterns. The extraction accuracy is estimated using the following indexes.

$$Precision = CDP/DDP$$
$$Recall = CDP/EDP$$
$$F\text{-}measure = 2 * Precision * Recall / (Precision + Recall)$$

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

We measured the scores of the above three indexes when modifying *thr* in the range 0.0 to 1.0 in increments of 0.2. Here, *thr* = 0 corresponds to the previous method, and *thr* > 0 indicates the proposed method that considers a causal relationship in a linkage pattern.

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

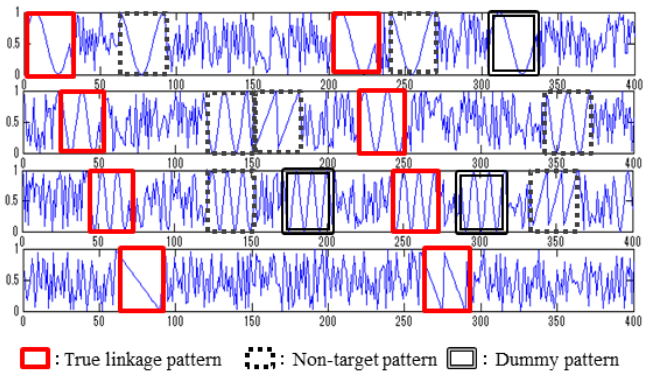

Fig. 3. Artificial dataset

TABLE I
DATASETS DETAILS

Dataset	# of kinds of true patterns	# of embedded true patterns	# of non-target patterns	# of dummy linkage patterns
test1	1	20	20	0
test2	2	10	20	0
test3	4	5	20	0
test4	2	10	20	0
test5	1	20	20	10
test6	4	5	20	10

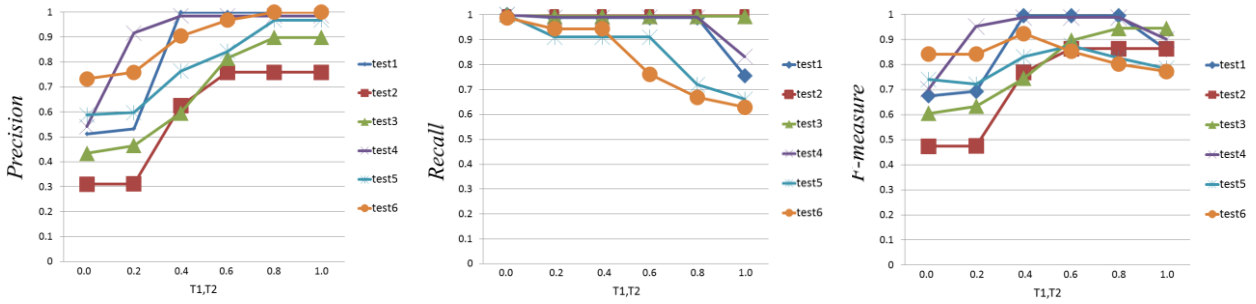

Fig. 4. Recommendation accuracy for the six datasets

IV. Results And Discussion

Figure 4 shows a graph of the *precision*, *recall*, and *F-measure*.

Precision increases with increasing *thr* for all datasets. In addition, the scores for the proposed method (*thr* > 0.0) are significantly greater than the previous method (*thr* = 0.0). This indicates that dummy linkage patterns can be eliminated by considering the causal relationship.

Recall shows approximately perfect scores in the range 0.0 to 0.8 for test1–test4. However, *recall* decreases with increasing *thr* for the test5 and test6 datasets because the completeness of the detection of true patterns decreases due to the strict conditions of the causal relationship.

F-measure is an index that considers *precision* and *recall*. From the results, *F-measure* decreases when *thr* is an extremely small or large value. For all datasets used in this experiment, we can see that *thr* should be set to approximately 0.6 to obtain stable and high extraction accuracy.

V. Conclusion

We have proposed a new linkage pattern mining method based on causal relationships among patterns appearing in multiple sequential data. We evaluated the extraction accuracy experimentally using artificial datasets. The results show that the proposed method can extract linkage patterns that demonstrate a causal relationship with high extraction accuracy compared to the previous method. In addition, we found that the parameter *thr* should be set to approximately 0.6 for stable and high extraction accuracy.

In the future, we will develop a more noise-robust method and apply it to real datasets.

References

[1] A. Achar, S. Laxman, and P. S. Sastry, "A Unified View of the Apriori-based algorithms for Frequent Episode Discovery," *in Knowledge and Information Systems (KAIS)*, vol. 31, no. 2, 2012, pp. 223–250.

[2] H. Xiong, P. Tan, V.Kumar, "Mining strong affinity association patterns in data sets with skewed support distribution," *in: Proceedings of the Third IEEE International Conference on Data Mining*, ICDM, 2003, pp. 387–394.

[3] H. Mannila, H. Toivonen, and A. I, Verkamo, "Discovery of frequent episodes in event sequences," *Data Mining and Knowledge Discovery*, vol. 1, 1997, pp. 259–289.

[4] K. Koperski and J. Han, "Discovery of spatial association rules in geographic information databases," In *Proc. 4th Int. Symp. Advances in Spatial Databases*, Vol. 951. Springer-Verlag, 1995, pp. 47–66.

[5] M. J. Zaki, "SPADE: An Efficient Algorithms for Mining Frequent Sequences," *Machine Learning*, Vol. 40, 2001, pp.31–60.

[6] C. Gong, W. Xindong, and Z. Xingquan, "Mining sequential patterns across time sequences," *New Generation Computing*, vol. 26, 2008, pp.75–96.

[7] S. Lee, T. Miura, Y. Okubo, and Y. Okada, "Linkage Pattern Mining Method for Multiple Sequential Data with Noise," *IAENG International journal of computer science*, vol.42, no.4, 2015, pp. 361–367.

[8] R. Miller and T. Yang, "Association Rules Over Interval Data," *In Proceedings of the 1997 ACM- SIGMOD Conference on Management of Data*, 1997, pp. 452–461.

[9] D. Apiletti, E. Baralis, and G. Bruno, T. Cerquitelli," Real-time analysis of physiological data to support medical applications," *Trans. Info. Tech. Biomed*, 2009, pp. 313–321.

[10] D. Sow, D. Turaga, M. Schmidt, "Mining of Sensor Data in Healthcare: A Survey," *In Managing and Mining Sensor Data*, 2013, pp. 459–504.

[11] A. Mannini, A. M. Sabatini, "Machine learning methods for

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

classifying human physical activity from on-body accelerometers," *Sensors*, 2010, pp. 1154–1175.

[12] F. Hu, M. Jiang, L. Celentano, and Y. Xiao, "Robust medical ad hoc sensor networks (MASN) with wavelet-based ECG data mining," *Ad Hoc Netw*, 2008, pp. 986–1012.

[13] H. Cheng, X. Yan, and J. Han, "Seqindex: Indexing sequences by sequential pattern analysis," *In Proc.2005 SIAM Int. Conf. Data Mining* (SDM'05), 2005, pp. 601–605.

[14] A. Negarestani, S. Setayeshi, M. Ghannadi-Maragheh, B. Akashe, "Layered neural networks based analysis of radon concentration and environmental parameters in earthquake prediction," *Journal of Environmental Radioactivity*, Volume 62, 2002, pp. 225–233.

[15] W. Dzwine et al, "Non multidimensional scaling and visualization of earth quake cluster over space and feature space," *nonlinear processes in geophysics 12*, 2005, pp.1–12.

[16] T. Miura and Y. Okada, "Detection of linkage patterns repeating across multiple sequential data," *International Journal of Computer Applications*, vol. 63, no. 3, 2013, pp. 14–17.

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)