Linkage Pattern Mining using Interval and Order of Pattern Appearance

Saerom Lee, Kaiji Sugimoto, and Yoshifumi Okada

***Abstract*— Linkage pattern mining is a data mining technique employed to extract a linkage pattern, that is, a set of frequent patterns appearing repeatedly across multiple sequential data. In this technique, when frequent patterns appear in the same time zone for multiple sequential data, they are extracted as a linkage pattern even if these patterns are neither correlated nor similar. Thus, this mining method is expected to become a promising approach for predicting the risks associated with disease and analysis of voice data. However, the existing linkage pattern mining method cannot extract the linkage pattern in which no overlap on the time axes is identified in its frequent patterns even if those frequent patterns obviously show a continuous appearance. In addition, there is another serious problem in the existing method; namely, in any two linkage patterns composed of the same frequent patterns, even if the order of frequent patterns for each other is different, these linkage patterns are mistakenly regarded as an identical linkage pattern. To solve these problems, we propose a new linkage pattern mining method that considers the interval and appearance order of the frequent patterns. Using artificial datasets, we further performed experiments to compare the extraction accuracies of the proposed and previous methods. The result shows that compared with the previous method, the proposed method allows the detection of linkage patterns correctly and comprehensively.**

***Index Terms*— sequential pattern mining, linkage pattern mining, appearance interval, appearance order**

I. Introduction

SEQUENTIAL pattern mining is a promising and effective data mining technique for obtaining useful information or knowledge from sequential data [1]. Hence, it has been applied to various fields such as prediction of disease from vital data and analysis of voice data [2]–[8]. However, the existing studies that aim at extracting similar or correlative patterns among multiple sequential data [9] mainly focus on sequential data derived from an identical domain. In recent years, large-scale sequential data have yielded in various domains [10], [11]. Thus, in the future, the discovering of useful information across big data in different domains will become an important technical task.

Manuscript received February 22, 2019. 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). K. Sugimoto is with the Division of Information and Electronic Engineering, Muroran Institute of Technology, 27-1, Mizumoto-cho, Muroran, Hokkaido 050-8585, Japan (e-mail: sugimoto@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: +81-143-46-5408; fax: +81-143-46-5408; e-mail: okada@csse.muroran-it.ac.jp).

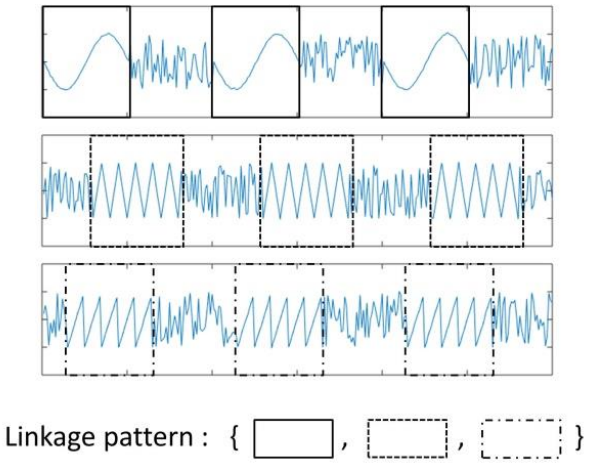

Fig. 1. Outline of the previous method

In our previous study, we proposed a linkage pattern mining method that targets multiple sequential data that have been derived from different domains [12]–[14]. Fig. 1 illustrates the outline of the previous method which extracts a set of frequent patterns that occur repeatedly across multiple sequential data in the same time zone. This set of frequent patterns is called a linkage pattern; it can be used to extract a meaningful set of patterns even if no similarity or correlation exists among frequent patterns in each sequential data. Nevertheless, there still exists a critical problem in this method; it is impossible to extract linkage pattern that is composed of the frequent patterns of which do not overlap in the same time zone but shows obviously a continuous appearance (hereafter referred to as a linked tendency). In addition, since this method extracts the linkage pattern based on the temporal overlap among frequent patterns, the appearance time for each frequent pattern is not considered. This causes a serious problem; the linkage patterns that comprise the same frequent patterns appearing in different orders are considered as identical linkage patterns.

To solve these problems, we propose a new linkage pattern mining technique based on the interval and order of frequent pattern appearance. Furthermore, the proposed method is used in extracting linkage patterns appearing not only across multiple sequential data but also in a single sequential data. To evaluate the extraction accuracy of the proposed method by comparing it with that of the previous method [14], we performed experiments using artificial sequential datasets.

The rest of this paper is organized as follows: The procedure and problems of the previous method are explained

in Section II. Section III explains the procedure of the proposed method. The details of the evaluation of the experimental performance using artificial sequential datasets are presented in Section IV. Furthermore, Section V presents the experimental results and discussions. Finally, an overall summary is presented in Section VI.

II. Previous Method

In this section, we present the procedure and problems of the previous method. The previous method employed an interval graph model to extract the linkage pattern, which represented an overlap of intervals of events that appeared in sequential data. In an interval graph, each interval is represented as a node and an edge between nodes represents an overlap between any two intervals [15]–[17]. In the previous method, each frequent pattern was represented as a node and temporal overlap between the frequent patterns in different sequential data was represented as an edge. If the number of the appearance of the interval graph is equal to or larger than a specified threshold, the previous method extracts a set of frequent patterns within the interval graph as a linkage pattern. However, two problems were identified in the previous method and they are as follows: In the first problem, if frequent patterns do not overlap the time axes as shown in Fig. 2(a), those patterns cannot be extracted as a linkage pattern despite showing linked tendencies. In the second problem, the previous method did not consider the appearance order of frequent patterns. For example, in Fig. 2(b), owing to the difference in their appearance orders, the two sets, including the frequent pattern X, Y, and Z, should be regarded as the different set; hence, no linkage pattern is extracted in this case. However, owing to the aforementioned reasons, it was observed that the pseudo linkage patterns were extracted.

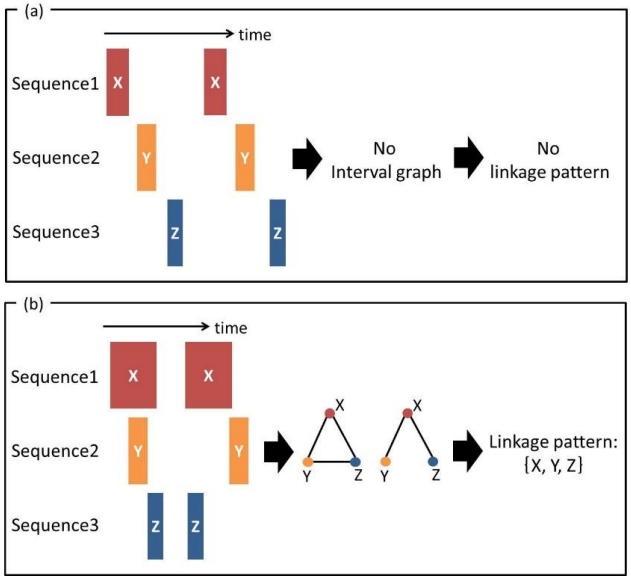

Fig. 2. Problems of the previous method

III. Proposed Method

In this study, we resolve the problems of the previous method by extracting the linkage pattern based on the appearance interval and appearance order of the frequent patterns. Fig. 3 shows the outline of the proposed method; it comprises the following five steps: 1) preprocessing (see Fig. 3a), 2) frequent pattern extraction and labeling (see Fig. 3b), 3) calculation of pattern width (see Fig. 3c), 4) merge of frequent patterns based on pattern width (see Fig. 3d), and 5) output of linkage pattern (see Fig. 3e). Step 3 (Fig. 3c), Step 4 (Fig. 3d), and Step 5 (Fig. 3e) are the newly added steps in this study. We describe the details of the aforementioned steps in Section A through Section E.

A. Preprocessing

In the proposed method, input data are comprised of multiple sequential data. First, each sequential data is normalized from 0 to 1. Next, this range is equally divided into d grades and each normalized data is assigned to any one of the grades.

B. Frequent pattern extraction and labeling

In this procedure, frequent patterns are first extracted from each sequential data using Mannila's algorithm [18], which uses two parameters, w and θ_F. w is the window width of the subsequence used in searching for frequent patterns from sequential data; θ_F is the minimum number of occurrences of the frequent patterns to be extracted.

Next, an identical label is given to the same frequent pattern as shown in Fig. 3(a). In this step, when the length of the frequent pattern is less than $w/2$, these frequent patterns are excluded without being labeled. If a frequent pattern is fragmented into several subsequences, the subsequence with the maximum length is labeled and the others are excluded.

C. Calculation of pattern width

In this section, we describe the calculation of the pattern width, which is an index used in judging whether any two frequent patterns demonstrate a linked tendency. First, two frequent patterns that appear first and second on the time axes are specified in all the sequential data. The set of frequent patterns that has the same label as the first pattern is named a reference set. The frequent pattern that appears at the i-th position (i = 1, 2,…, n) in the reference set is denoted as R_i. The set of frequent patterns that has the same label with the second pattern is named a target set. The frequent pattern that appears at the j-th position (j = 1, 2,…, m) in the target set is denoted as T_j. In the pattern width, pw_{ij} is defined as the difference between the start time of R_i and the end time of T_j. This value is calculated for all the combinations of R_i and T_j. In Fig. 3(b), the sets of frequent patterns labeled X and Y are regarded as the reference set and target set, respectively.

D. Merge of frequent patterns based on pattern width

Here, we describe the method for merging frequent patterns to show linked tendency based on the pattern width. First, we calculate the median and the interquartile range (IQR) of a set of the pattern widths. The IQR formula is represented by the difference of the first quartile Q1 from the third quartile Q3 given as follows:

$$\mathrm{IQR} = (\mathrm{Q3} - \mathrm{Q1}).$$

Next, we identify a set of an ordered pair (R_i, T_j) that expresses a combination of R_i and T_j, the pattern width of which is within the range of the median ± IQR/p. Here p is a

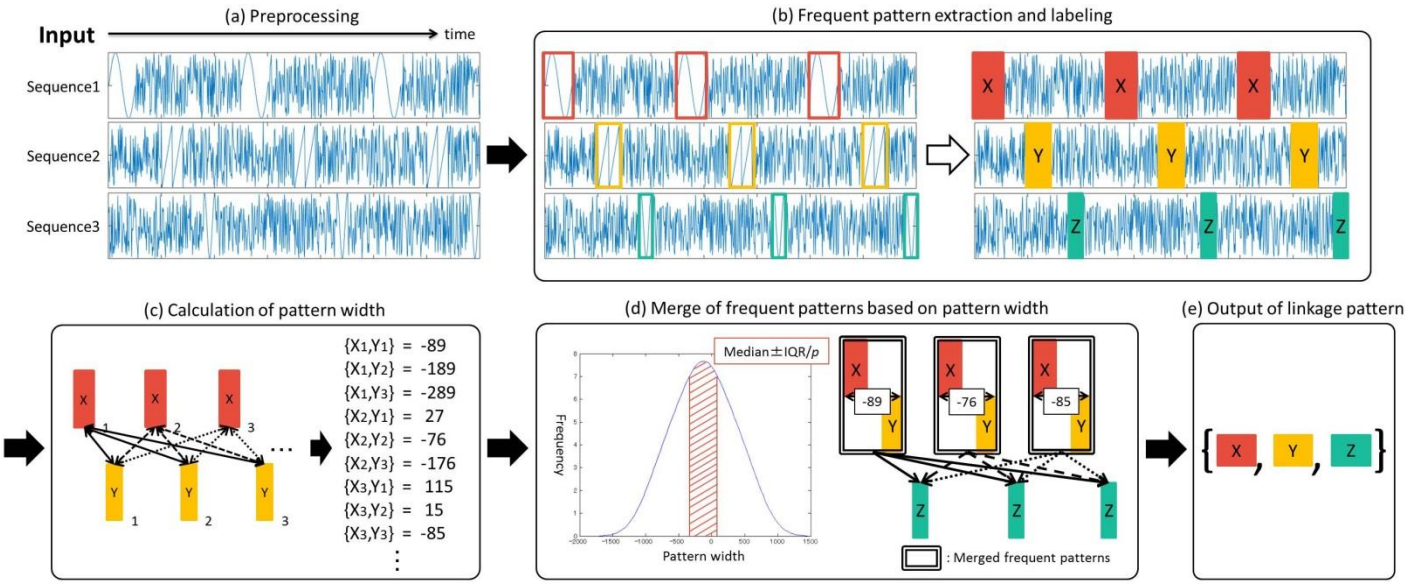

Fig. 3. Procedure of the proposed method

parameter used to adjust the error of the pattern width. If the size of the aforementioned set is equal to or greater than θ_L, R_i and T_j that satisfy this condition are merged and regarded as one frequent pattern.

E. The output of linkage pattern

The processes in Sections C and D are repeated for all frequent patterns until no frequent pattern is merged, and the remaining frequent patterns are output as linkage patterns. Furthermore, as revealed in Fig. 3, the pattern Z satisfies the condition for merging as discussed in Section D. Thus, a set of patterns X, Y, and Z is extracted as a linkage pattern. Fig. 4 shows the pseudo code for the algorithms described in Section C through Section E.

IV. Experiments

We applied the previous and proposed methods to artificial datasets and evaluated their extraction accuracies of linkage patterns using visual inspection and evaluation indexes.

A. Artificial datasets

We used seven artificial datasets (Dataset 1–Dataset 7) as shown in Fig. 5. Each dataset was created by embedding artificial linkage patterns into random sequential data generated from uniform random numbers. Dataset 1–Dataset 6 include a linkage pattern surrounded by the solid frame and Dataset 7 includes three linkage patterns surrounded by the solid, dashed, and dashed dotted frames. In addition, Datasets 2, 5, 6, and 7 include frequent patterns that do not belong to any linkage pattern, which are surrounded by the dotted frame.

B. Parameter settings

In the previous and proposed methods, the parameters d, w, and θ_L were set to 50, 5, and 3, respectively, and θ_F was set to 3, 4, or 5. In the previous work, these values showed the best results of the extraction accuracies for each artificial dataset [14]. However, the proposed method additionally requires a new parameter p for adjusting the error of pattern width. The parameter p was set to 5, 10, 15, 20, 25, 30, 35, 40, 45, and 50.

Algorithm 1 Proposed method Step C, D, and E

Input: $R_i(i = 1, 2, ..., n)$: Pattern of reference set;
R : A set of reference patterns; (The set of frequent patterns with a same label as the first pattern is named a reference set)
$T_j(j = 1, 2, ..., m)$: Pattern of target set;
T : A set of target patterns; (The set of frequent patterns with a same label as the second pattern is named a target set)
(R_i, T_j) : Ordered pair of R_i and T_j;
O : A set of ordered pair of R_i and T_j;
θ_L : The minimum number of occurrences of frequent patterns to be extracted;
Output: Linkage pattern

```
main
if |R| ≠ 0 then
    while |T| ≥ 1 do
        Function PATTERNWIDTH()
        Function MERGE()
    end while
    Function OUTPUT()
end if
```

Function PATTERNWIDTH

```
for i = 1 to n do
    for j = 1 to m do
        PW_ij = The start time of R_i − The end time of T_j
    end for
end for
```

Function MERGE

```
for i = 1 to n do
    for j = 1 to m do
        if Median − IQR/p ≤ PW_ij ≤ Median + IQR/p then
            O+ = (R_i, T_j)
        end if
    end for
end for
if |O| ≥ θ_L then
    R = 0
    T = 0
    for k = 1 to |O| do
        R_k = new pattern(O_k)
    end for
else
    T is masked        ▷ Remove the T from the linkage pattern candidate
end if
T = Frequent patterns with a same label as the second pattern
O = 0
```

Function OUTPUT

```
Linkage pattern = R
```

Fig. 4. Pseudo code for the algorithms of step C, D, and E

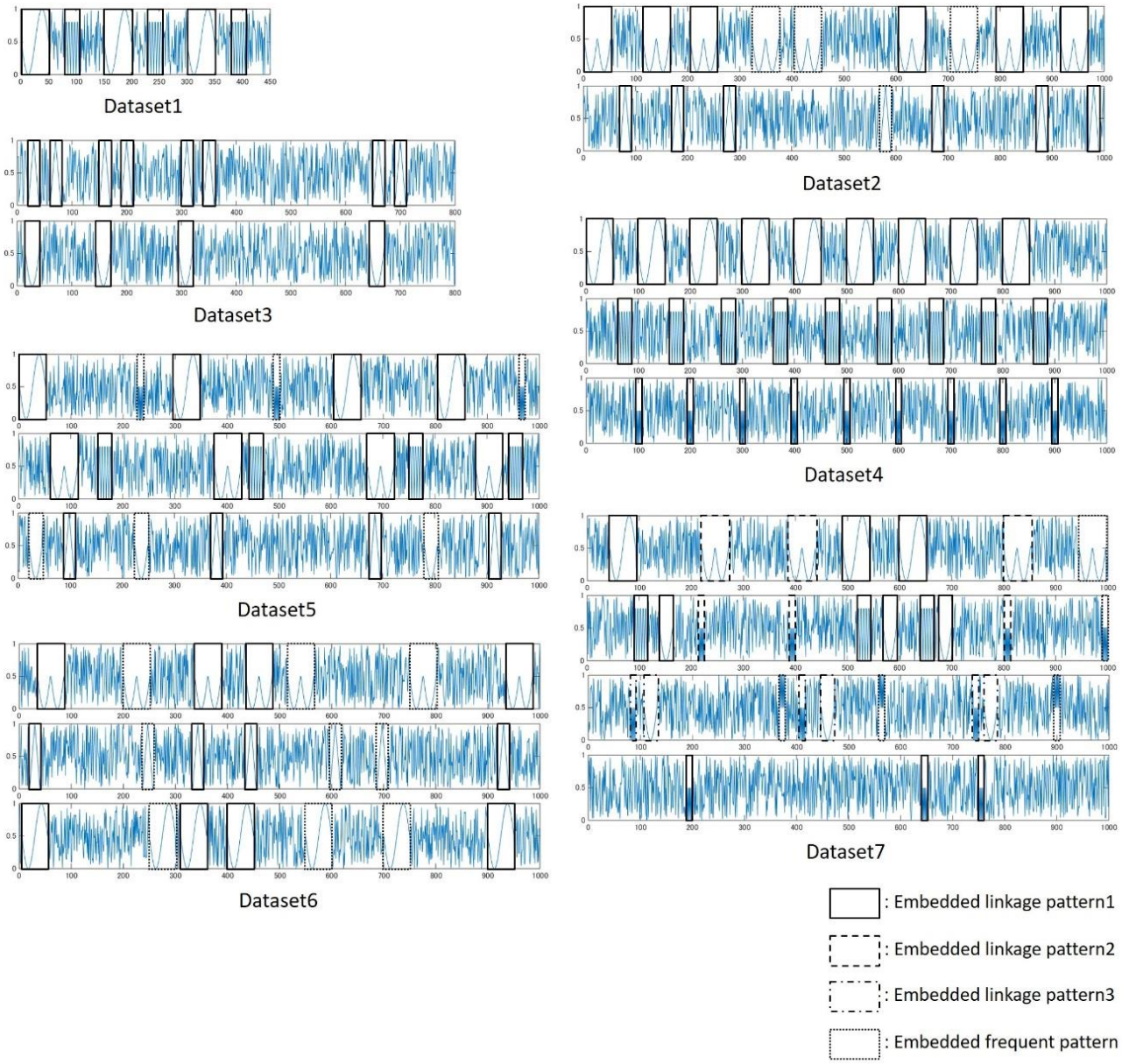

Fig. 5. Artificial datasets

C. *Extraction accuracy of the linkage patterns*

In this study, precision and recall were used as the evaluation indexes. These indexes were calculated as follows:

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

Here, CDP is the number of data points in the correct detected areas of the embedded linkage patterns, whereas DDP is the number of data points in the areas of the embedded linkage patterns detected by the previous and proposed methods. EDP is the number of data points in the embedded linkage patterns.

V. Results and Discussion

A. *Visualization result*

Fig. 6 shows the visualization results of the linkage patterns extracted from their respective artificial datasets, showing the best results of the previous and proposed methods. The colored parts show the extracted linkage patterns and their colors correspond to those of the frames shown in Fig. 5. In the previous method, the embedded linkage patterns were not extracted from Dataset 1 and Dataset 2; therefore, the results were not shown in Fig. 6. From these results, the previous method only extracts the linkage pattern in which the appearances of the frequent patterns overlap on the time axes. However, the proposed method can adequately extract the linkage patterns in which the frequent pattern shows linked tendency regardless of whether these overlap on the time axes. In Datasets 6 and 7, we can see that the previous method regards the set of the frequent patterns that appear in different orders as the linkage pattern. In contrast, the proposed method can extract only the linkage patterns whose frequent patterns appear in the same order, this is mainly because the proposed method searches for linkage patterns by checking the appearance order of their frequent patterns. Based on the results of Datasets 1, 3, 5, and 7, we observe that the proposed method can extract the linkage patterns that appear in a single sequential data. This is achieved because the proposed

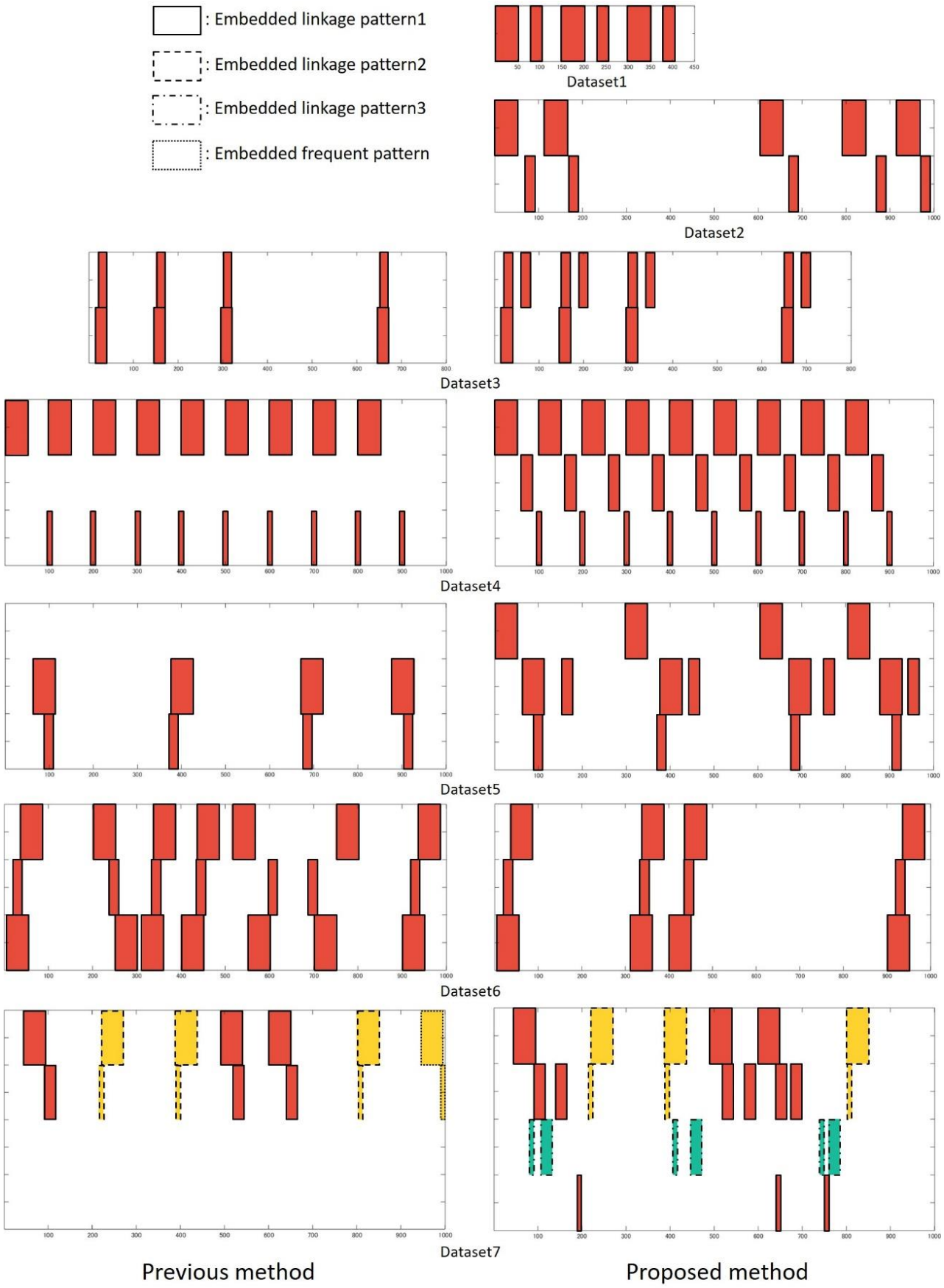

Fig. 6. Visualization results

method searches linkage patterns by checking pattern widths of frequent patterns in each sequential data and among different sequential data.

B. Extraction accuracy

Fig. 7 shows comparison results of the extraction accuracies between the previous and proposed methods; thus,

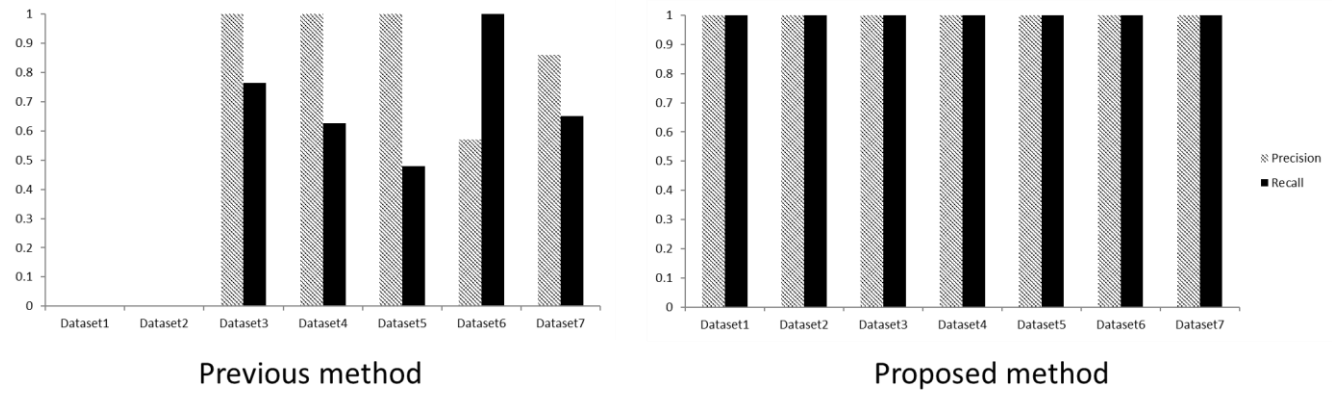

Fig. 7. Comparison results of the extraction accuracies between the previous and proposed method

revealing the best results in the previous and proposed methods. As stated in the previous section, no linkage patterns were found in Dataset 1 and Dataset 2; therefore, these scores were not displayed in Fig. 7. From the results, it is revealed that the proposed method demonstrates perfect scores in all the artificial datasets. In contrast, the scores of the previous method are lower than those of the proposed method and change considerably depending on the datasets. As revealed from these results, compared with the previous method, the proposed method assists in detecting linkage patterns correctly and comprehensively.

C. Impact of the parameter p

Because the parameter *p* is used to determine whether any two frequent patterns show a linked tendency, it is considered to have a big impact on extraction accuracy. Hence, we investigated the impact of parameter *p* on extraction accuracy. Fig. 8 shows the precision and recall of different *p*. From these figures, we can see that precision tends to decrease in extremely small *p*, whereas the recall tends to decrease in extremely large *p*. When *p* is small, the pattern width becomes wider. This means that irrelevant frequent patterns may be mistakenly merged because it causes a decrease in the precision score. When *p* is large, the pattern width becomes narrower, implying that the relevant frequent patterns may not be merged; hence it causes a decrease in the recall score. Overall, we found that it is reasonable to set the parameter *p* to approximately 20.

D. Comprehensive evaluations of the proposed method

So far, many pattern mining algorithms for multiple sequential data have been proposed [19-26]. These algorithms focused on single kind of sequential data, i.e., sequential data derived from an identical domain. On the other hand, there are several studies targeting different domains [27, 28]. These algorithms aimed at extracting unique sequence patterns in each domain.

In the above studies, extended sequential pattern algorithm [21] has been proposed and its usefulness has been shown. However, these methods only extract frequent patterns that repeatedly appear in each sequential data and do not provide a function to automatically detect their relevance or rules.

Conversely, our method can extract not only frequent patterns from each sequential data of different domains but also linkage patterns that are sets of frequent patterns related to each other across multiple sequential data. In recent years, different types of big data have been produced from various fields. A technique for extracting relationships across such different domains will become a promising tool to obtain useful cross-domain information. Our method can be used to detect the relationships and rules between different types of sequential data, e.g., different physiological data such as electrocardiograms and electroencephalograms; hence it will provide informative findings for disease diagnosis or emotion recognition.

The advantages of the proposed method are as follows:

- It is possible to detect relations between different domains.
- It is possible to extract linkage patterns consisting of the frequent patterns that have no overlap on the time axes across multiple sequential data.
- It is possible to extract linkage patterns considering the order of frequent patterns.

The limitations of the proposed method are as follows:

- It requires much computational time because of the combinatorial search in the frequent pattern mining based on Mannila's algorithm.
- It requires much computational time because of the combinatorial checking for frequent patterns composing the linkage pattern.

To solve the above problems, we need further improvement of the algorithm.

VI. CONCLUSIONS

In this study, we proposed a new linkage pattern mining method based on the interval and order of appearance of frequent patterns. From the experimental results, we used seven artificial datasets to reveal that the proposed method outperformed the previous method as it can extract the following three types of linkage patterns that were not extracted by the previous method:

1) Linkage pattern appearing in a single sequential data;

2) Linkage pattern comprising the frequent patterns with no overlap on the time axes across multiple sequential data;

3) Linkage pattern in which the frequent patterns appear in an identical order on the time axes.

Furthermore, it was found that it is reasonable to set the parameter *p* to around 20. In the future study, we will develop a method for automatically tuning the parameter *p* based on the distribution of the pattern width. In addition, we will apply the proposed method to real datasets, such as

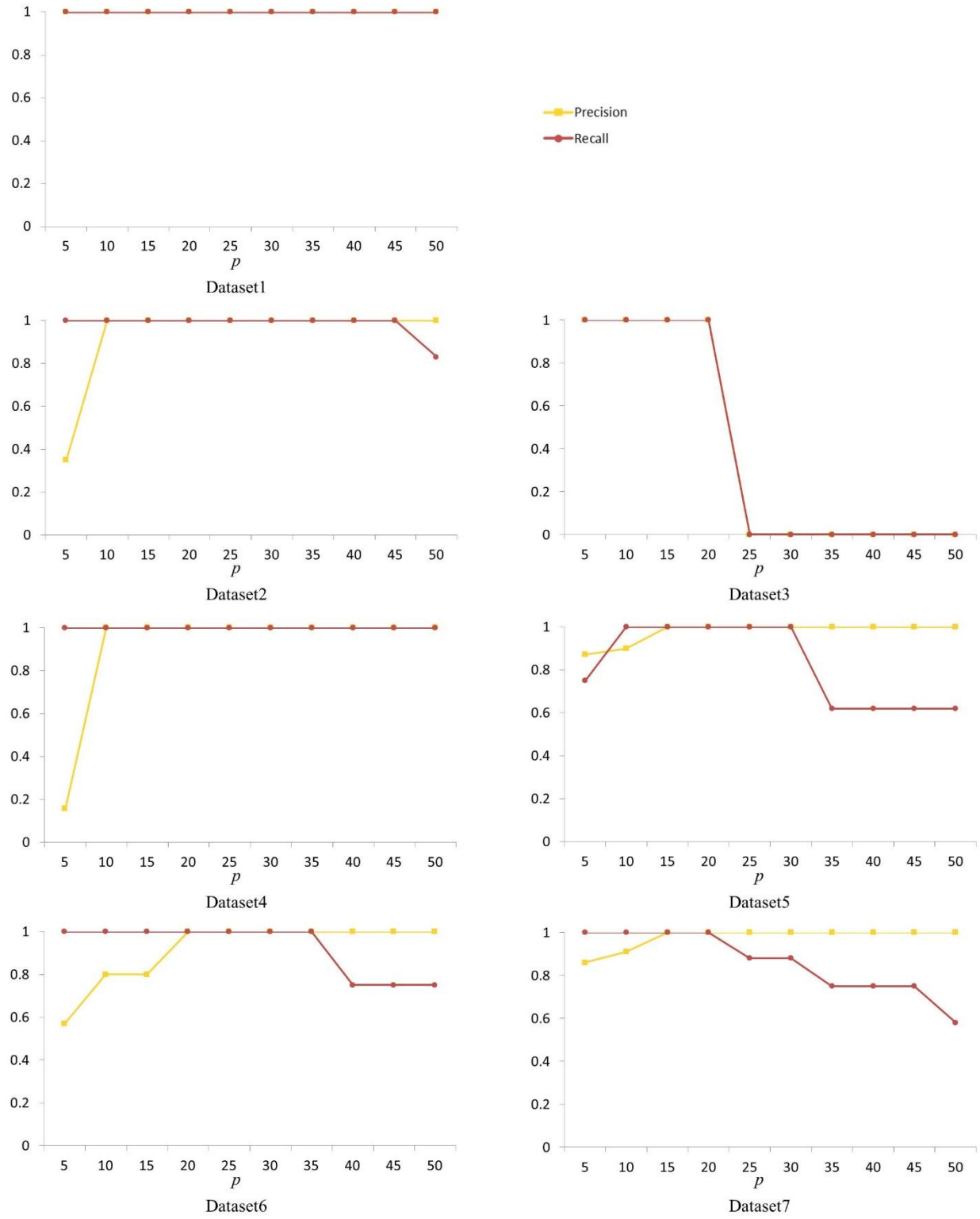

Fig. 8. Extraction accuracies of the proposed method in different *p*

electrocardiogram data and voice data.

ACKNOWLEDGMENT

This work was supported by Grant-in-Aid for Scientific Research (C) (No. 17K00373) from the Japan Society for the Promotion of Science.

REFERENCES

[1] R. Agrawal and R. Srikant, "Mining sequential patterns," in *1995 Proc. 11th Int. Conf. on Data Engineering*, pp. 3–14, 1995.

[2] F. Takchungm, "A review on time series data mining," *Eng. Appl. Artif. Intell.*, vol. 24, no. 1, pp. 164–181, 2011.

[3] Q. Zhao and S. S. Bhowmick, "Sequential pattern mining: A survey technical report." *CAIS*, Nanyang Technological University, Singapore, No. 2003118, 2003.

[4] C. I. Ezeife and Y. Lu, "Mining web log sequential patterns with position coded pre-order linked WAP-tree," *Data Min. Knowl. Discov. Springer Science, Business Media. Inc. Manufactured in The Netherlands*, vol. 10, pp. 5–38, 2005.

[5] X. Wu, Y. Wu, Y. Wang, and Y. Li, "Privacy-aware market basket data set generation: A feasible approach for inverse frequent set mining," *in 2005 Proc. 5th SLAM Int. Conf. on Data Mining*, pp. 103–114, 2005.

[6] A. D. Lattner, A. Miene, U. Visser, and O. Herzog, "Sequential pattern mining for situation and behavior prediction in simulated robotic soccer," *RoboCup 2005: Robot Soccer World Cup IX Lecture Notes in Computer Science*, vol. 4020, pp. 118–129, 2006.
[7] R. Sarno, R. D. Dewandono, T. Ahmad, M. F. Naufal, and F. Sinaga, "Hybrid association rule learning and process mining for fraud detection," *IAENG International Journal of Computer Science*, vol. 42, no. 2, pp. 59–72, 2015.
[8] M. Karaca, M. Bilgen, A. N. Onus, A. G. Ince, and S. Y. Elmasulu, "Exact tandem repeats analyzer (E-TRA): A new program for DNA sequence mining," *J. Genet.*, vol. 84, pp. 49–54, 2005.
[9] Md. M. Monwar, and S. Rezaei, "An efficient parallel processing approach for multiple biological sequence alignment," *IAENG International Journal of Computer Science*, vol. 33, no. 2, pp. 32–36, 2007.
[10] T. W. Liao, "Clustering of time series data—a survey," *Pattern Recognit.*, vol. 38, no 11, pp. 1857–1874, 2005.
[11] S. Fujii, K. Kudo, T. Ohtsuki, and S. Oda, "Tapping performance and underlying wrist muscle activity of nondrummers, drummers, and the world's fastest drummer," *Neurosci Lett*, 459, pp. 69–73, 2009.
[12] T. Miura and Y. Okada, "Detection of linkage patterns repeating across multiple sequential data," *Int. J. Computer Applications*, vol. 63, no. 3, pp. 14–17, 2013.
[13] S. Lee, T. Miura, and Y. Okada, "A new method for improving the performance of linkage pattern mining," *Proc. of IMECS*, pp. 36–40, 2014.
[14] 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, pp. 361–367, 2015.
[15] N. Miyoshi, T. Shigezumi, R. Uehara, and O. Watanabe, "Scale free interval graphs," *Theor. Comput. Sci.*, vol. 410, no. 45, pp. 4588–4600, 2009.
[16] N. Korte and R. H. Mohring, "An incremental linear-time algorithm for recognizing interval graphs," *SIAM J. Computing*, vol. 18, pp. 68–81, 1979.
[17] G. S. Lueker and K. S. Booth, "A linear time algorithm for deciding interval graph isomorphism." *J. ACM*, vol. 26, pp. 183–195, 1979.
[18] H. Mannila, H. Toivonen, and A. I. Verkamo, "Discovery of frequent episodes in event sequences," *Data Min. Knowl. Discov.*, vol. 1, pp. 259–289, 1997.
[19] G. Chen, X. Wu, and X. Zhu, "Sequential Pattern Mining in Multiple Streams," *IEEE Intl. Conf. on Data Mining (ICDM'05)*, 4 pp.–, 2005
[20] D. Töws, M. Hassani, C. Beecks, and T. Seidl, "Optimizing sequential pattern mining within multiple streams," *Database Syst. for Bus., Technol. and Web (BTW 2015)*, vol. P-242, pp. 223–232, 2015.
[21] J. Han, J. Pei, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M.C. Hsu, "Prefixspan: mining sequential patterns efficiently by prefix-projected pattern growth," in *Procs. 17th International Conference on Data Engineering*, pp. 215–224, 2001.
[22] H.C. Kum, J.H. Chang, and W. Wang, "Sequential pattern mining in multi-databases via multiple alignment," *Data Min. Knowl. Discov.*, vol. 12, no. (2–3), pp. 151–180, 2006.
[23] P. Fournier-Viger, A. Gomariz, M. Campos, and R. Thomas, "Fast vertical mining of sequential patterns using co-occurrence information," *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 40–52, 2014.
[24] M.J. Zaki, "SPADE: an efficient algorithm for mining frequent sequences," *Kluwer Academic Publishers. Manufactured in The Netherlands. Machine Learnin*g, vol. 42 pp. 31–60, 2001.
[25] J. Ayres, J. Flannick, J. Gehrke, and T. Yiu, "Sequential pattern mining using a bitmap representation," *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 429–435, 2002.
[26] A. Gomariz, M. Campos, R. Marin, and B. Goethals, "ClaSP: an efficient algorithm for mining frequent closed sequences," *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 50–61, 2013.
[27] W. Peng, and Z. Liao, "Mining sequential patterns across multiple sequence databases," *Data Knowl.*, vol. 68, pp. 1014–1033, 2009.
[28] M. Hassani, C. Beecks, D. Töws, T. Serbina, M. Haberstroh, P. Niemietz, S. Jeschke, S. Neumann and T. Seidl, "Sequential pattern mining of multimodal streams in the humanities," *Database Syst. for Bus., Technol. and Web (BTW 2015)*, pp. 683–686, 2015.