Letter

Pattern Matching of Industrial Alarm Floods Using Word Embedding and Dynamic Time Warping

Wenkai Hu, Xiangxiang Zhang, Jiandong Wang, Guang Yang, and Yuxin Cai

Dear Editor,

This letter proposes a new pattern matching method based on word embedding and dynamic time warping (DTW) to identify groups of similar alarm floods. First, alarm messages are transformed into numeric values that represent alarms and also reflect the relationships between alarm occurrences. Then, similarities between numerically encoded alarm flood sequences are calculated by DTW and groups of similar floods are identified via clustering. The effectiveness of the proposed method is demonstrated by a case study with alarm & event data obtained from a public industrial simulation model.

In large-scale industrial facilities, alarm overloading and alarm flood issues are major problems compromising the performance of alarm systems; especially, the presence of alarm floods can make an alarm system partially or completely fail and thus lead to serious consequences [1]. In the ANSI/ISA-18.2 standard [2], an alarm flood is known as a situation that the amount of alarms exceeds the management capability of an operator. Motivated by the fact that alarm floods in historical alarm data may resemble each other and hold certain common subsequences that imply same underlying root causes, pattern matching techniques have been proposed to discover the recurring patterns among alarm flood sequences.

Existing approaches for alarm flood pattern matching are mostly based on biological sequence alignment, such as the Smith-Waterman (SW) and Needleman-Wunsch (NW) algorithms [3]. In [4], a modified SW algorithm was proposed to detect local alignments between alarm floods and tolerate the order ambiguity for alarms that occur almost simultaneously. To improve the computational efficiency, an accelerated alarm flood sequence alignment method was proposed in [5]. Toward early warning of alarm floods, a real-time pattern matching method was developed in [6] to provide a ranking list of similar historical floods for operators during online monitoring. Reference [7] formulated a generalized pattern matching approach for comparison of alarm floods generated from several individual facilities that are similar in architecture and functionality.

The above studies provide effective solutions for industrial practitioners to identify similar alarm floods. It is noteworthy that the biological sequence alignment algorithms are specifically designed to compare strings of nucleic acid or protein sequences. Despite that there are differences among the alarm flood pattern matching methods in literature, they are essentially all based on the SW or NW algorithms, and can only conduct similarity measure based on match operations for alarms represented by textual strings. As a result, the relationships between alarm occurrences are neglected in the match operations and thus may lead to erroneous conclusions. Accordingly, this letter proposes a new pattern matching method based on word embedding and DTW to reveal sequence similarities between alarm floods from a new perspective. To demonstrate the effectiveness of the proposed method, a case study is presented to make comparisons with classical biological sequence alignment approaches based on the alarm & event data obtained from a public industrial simulation model.

It is worth mentioning that [7] also adopted word processing as a prerequisite step to process alarm representations and then conducted pattern matching. There are three major differences between [7] and this letter: 1) Reference [7] targets at comparing alarm floods generated from several individual facilities that are similar in architecture and functionality, whereas the purpose of this study is to improve the accuracy of pattern matching by taking into account the relationships between alarm occurrences; 2) The word processing in [7] is used to distill key words from alarm messages and reconstruct alarm descriptions such that alarms of the same type can be represented equally, whereas the word embedding in this study transforms each alarm into a numeric vector whose value can reflect the relationships between alarm occurrences; 3) The generalized alarm representations are taken as individual strings in [7] and the SW algorithm is exploited for alarm flood pattern matching, whereas this study can conduct arithmetic calculation for alarms in numeric forms and thus takes the DTW algorithm for pattern matching.

Problem description: Denote an alarm & event (A&E) log $\mathbb{D}$ collected over a certain historical time period by $\mathbb{D} = \{a_j \in \mathcal{A}, j = 1, 2, \ldots, |\mathbb{D}|\}$, where $\mathcal{A}$ stands for the set of unique alarms in the alarm system, $\mathbb{D}$ indicates the size of the A&E log $\mathbb{D}$, and j is the order of the alarm event a_j in $\mathbb{D}$. Then, alarm floods can be identified by calculating the alarm rates in a window of 10 min and comparing with the benchmark thresholds in [2]. The extraction process of alarm floods can be found in [8]. An alarm flood is represented as a sequence consisting of chronologically ordered alarms, i.e.,

$$F = \langle a_1, a_2, \ldots, a_m \rangle \tag{1}$$

where $\langle \cdot \rangle$ represents a sequence, a_i denotes the ith alarm event in F, and m indicates the length of F.

Given a pair of flood sequences $F_x = \langle a_1^x, a_2^x, \ldots, a_m^x \rangle$ and $F_y = \langle a_1^y, a_2^y, \ldots, a_n^y \rangle$, the objective is to correctly measure the similarity S_{xy} between the two sequences, such that similar alarm floods can be identified and grouped. The most widely used pattern matching methods are based on biological sequence alignment, such as the SW and NW algorithms, which take alarm messages as strings in calculation and only the match operation is applied. Accordingly, alarms are independent from each other in the similarity analysis and the relationships between alarm occurrences are neglected, which may lead to erroneous conclusions. In practice, alarms may hold certain relations. For instance, the distance between two related alarms (e.g., "FI028.PVLO" and "FI028.LOLO") should be smaller than that between two unrelated ones (e.g., "FI028.PVLO" and "PC999. PVBAD"). Therefore, in similarity analysis, it would be more proper to measure the distances between alarms based on not only the alarm tags but also their relationships in occurrences.

Motivated by the above problem, this work proposes a new framework for pattern matching of industrial alarm floods. The proposed method includes two major steps: First, the textual alarm messages are transformed into numeric vectors by word embedding, such that each vector represents an unique alarm and its value can reflect the relationships between alarm occurrences; then, similarities between numerically encoded alarm flood sequences are calculated by DTW

Corresponding author: Jiandong Wang.

Citation: W. K. Hu, X. X. Zhang, J. D. Wang, G. Yang, and Y. X. Cai, "Pattern matching of industrial alarm floods using word embedding and dynamic time warping," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 4, pp. 1096–1098, Apr. 2023.

W. K. Hu, X. X. Zhang, G. Yang, and Y. X. Cai are with the School of Automation, China University of Geosciences, Wuhan 430074, Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan 430074, and with Engineering Research Center of Intelligent Technology for Geo-Exploration, Ministry of Education, Wuhan 430074, China (e-mail: wenkaihu@cug.edu.cn; zhangxiang2020@foxmail.com; Guang Yangc@163.com; 869507365@qq.com).

J. D. Wang is with the College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao 266590, China (e-mail: jiandong@sdust.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2023.123594

and groups of similar alarm floods are identified via clustering.

Alarm coding by word embedding: The purpose of alarm coding is to convert textual alarm messages into numeric vectors. Here, the Word2Vec, as a two-layer neural network used to produce word embedding [9] and [10], is exploited. The input is the whole corpus (namely, an A&E log) and the output is the set of word vectors that represent unique alarms. In view of the sensitivity to low frequency words, the skip-gram model in Word2Vec is applied. The training objective is to find word vector representations of alarms that help predict background words (namely, alarms) in the A&E log.

Given an A&E log $\mathbb{D}$ with N unique alarms, the kth alarm $a_k \in \mathcal{A}$ is represented as a binary vector η_k, i.e.,

$$\eta_k = [0,\ldots,0,1,0,\ldots,0]^T \tag{2}$$

where the length of η_k is N, the kth value of η_k is 1, and others are 0's. Next, each alarm is embedded into a M-dimensional real value space based on a weighting matrix $V \in \mathbb{R}^{N\times M}$. The structured representation vector $\mathbf{v}_k$ of $a_k \in \mathcal{A}$ is given by

$$\mathbf{v}_k = V^T \eta_k. \tag{3}$$

The objective of training a Skip-gram model is to maximize the average log probability Φ given by [9]

$$\max \Phi = \max\left(\frac{1}{|\mathbb{D}|}\prod_{i=1}^{|\mathbb{D}|}\prod_{-c\le j\le c, j\neq 0} P\left(a_{i+j} \mid a_i\right)\right) \tag{4}$$

where c is the context window size for the i th alarm a_i in $\mathbb{D}$; $P(a_{i+j} \mid a_i)$ is a conditional probability obtained by

$$P\left(a_{i+j} \mid a_i\right) = \frac{e^{\left(\mathbf{v}_{i+j}^T \mathbf{v}_i\right)}}{\sum_{k=1}^{N} e^{\left(\mathbf{v}_k^T \mathbf{v}_i\right)}}. \tag{5}$$

The optimization objective is further transformed to

$$\max \log \Phi = \max\left(\frac{1}{|\mathbb{D}|}\sum_{t=1}^{|\mathbb{D}|}\sum_{-c\le j\le c, j\neq 0} \log P\left(a_{i+j} \mid a_i\right)\right). \tag{6}$$

The word vector $\mathbf{v}_i$ can be obtained iteratively based on the update equation with respect to the derivative of the above objective function [10]. Then, $\mathbf{v}_i$ is utilized to encode the alarm a_i. As a result, an alarm flood sequence $F = \langle a_1, a_2, \ldots, a_m\rangle$ is transferred to the encoded form represented by

$$\tilde{F} = \langle \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m\rangle. \tag{7}$$

There are two key parameters in alarm encoding, namely, the dimension M of word vectors and the context window size c. A larger dimension M may better represent the relationships between alarms, but would lead to high computational cost. Analogously, a large window size c may make training of word vectors more accurate, but would increase the time of training. Thus, the two parameters should be properly set to ensure accuracy and computational efficiency.

Alarm flood similarity analysis based on DTW: Given two alarm flood sequences F_x and F_y, their encoded vector forms are represented by $\tilde{F}_x = \langle \mathbf{v}_1^x, \mathbf{v}_2^x, \ldots, \mathbf{v}_m^x\rangle$ and $\tilde{F}_y = \langle \mathbf{v}_1^y, \mathbf{v}_2^y, \ldots, \mathbf{v}_n^y\rangle$. Here, the DTW algorithm, as a commonly used method for similarity analysis of temporal sequences [11], is utilized to measure the similarity score S_{xy} between the encoded sequences. The reasons for choosing the DTW algorithm are as follows: 1) DTW is applicable to comparison between two temporal sequences in different lengths and thus fits the application in this study; 2) DTW carries out similarity calculation based on distance functions (e.g., Euclidean distance) for numeric data, making it directly applicable to the encoded alarm floods $\tilde{F}_x$ and $\tilde{F}_y$. Specifically, the alarm flood similarity analysis consists of four steps:

1) Calculation of distance matrix: A distance matrix M of dimension $m\times n$ is created for F_x and F_y. Each element $M(i,j)$ represents the distance between the ith alarm a_i^x in F_x and jth alarm a_j^y in F_y; $M(i,j)$ can be measured by the Euclidean distance based on the encoded word vectors $\mathbf{v}_i^x$ and $\mathbf{v}_j^y$ of a_i^x and a_j^y, i.e.,

$$M(i,j) = (\mathbf{v}_j^y - \mathbf{v}_i^x)^T (\mathbf{v}_j^y - \mathbf{v}_i^x). \tag{8}$$

2) Dynamic path warping: Further, the dynamic warping path is searched based on the distance matrix M. The shortest warping path is selected as the optimal warping path, which is found by calculating a cumulative distance matrix L with each element $L_{i,j}$ given by

$$L(i,j) = M(i,j) + \min\{L(i-1,j), L(i,j-1), L(i-1,j-1)\}. \tag{9}$$

The DTW distance between two alarm flood sequences F_x and F_y is then obtained as $L_{xy} = L(m,n)$.

3) Normalization of distance measure: In order to unify the scale of distances, it is necessary to normalize the calculated DTW distance and convert it into a similarity within $[0,1]$. A standard distance for sequences of different lengths is given by

$$\text{Dist}_{xy} = \frac{L_{xy}}{\sqrt{m^2+n^2}}. \tag{10}$$

Further, a normalized similarity score S_{xy} is calculated as

$$S_{xy} = \frac{\alpha \text{Dist}_{\max} - \text{Dist}_{xy}}{\alpha \text{Dist}_{\max}} \tag{11}$$

where $\text{Dist}_{\max}$ is the largest standard distance for different pairs of floods; α is a user defined value to ensure S_{xy} neither reach zero for the the largest standard distance, nor become too large for all alarm flood pairs. A reasonable choice is $\alpha \in [1,2]$.

4) Clustering of alarm floods: Given a set of H sequences F_h, $h = 1,2,\ldots,H$, the similarity score between each pair of sequences is calculated and accordingly a similarity matrix $S \in \mathbb{R}^{H\times H}$ is obtained. To identify the groups of similar floods, the spectral clustering is applied. It tries to learn a low-dimensional representation $U \in \mathbb{R}^{n\times g}$ (g is the number of clusters) [12] by solving

$$\min_{U^T U = I} \text{Tr}\left(U^T \lambda U\right) \tag{12}$$

where $\text{Tr}(\cdot)$ denotes the trace operation, I is the identity matrix, and $\lambda \in \mathbb{R}^{n\times n}$ is the Laplacian graph computed from $S \in \mathbb{R}^{H\times H}$. Eventually, alarm flood sequences are clustered into g groups; within each group, alarm floods hold high similarities with each other.

Case study: This section provides a case study to illustrate the proposed method. The vinyl acetate monomer (VAM) public model is used to simulate an alarm system and generate A&E data [13]. There are 130 alarm flood sequences extracted from the produced A&E data. These floods are associated with 13 different faults, and each fault is related to 10 floods. The lengths of alarm floods vary, ranging from 156 to 1108. The proposed pattern matching method is applied and compared with the SW and NW algorithms [3] and [4].

As aforementioned, the dimension M of word vectors and the context window size c may affect the pattern matching results. Hereby, simulations are conducted to investigate the influence of the two parameters. Both M and c are set to change from 1 to 10. Fig. 1 presents the average misclassification rate versus one parameter with

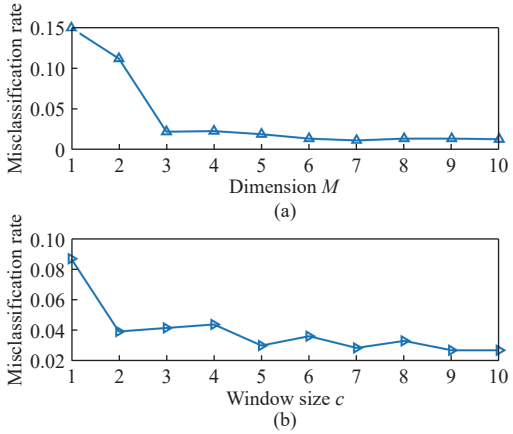

Fig. 1. Effects of setting different values of M and c.

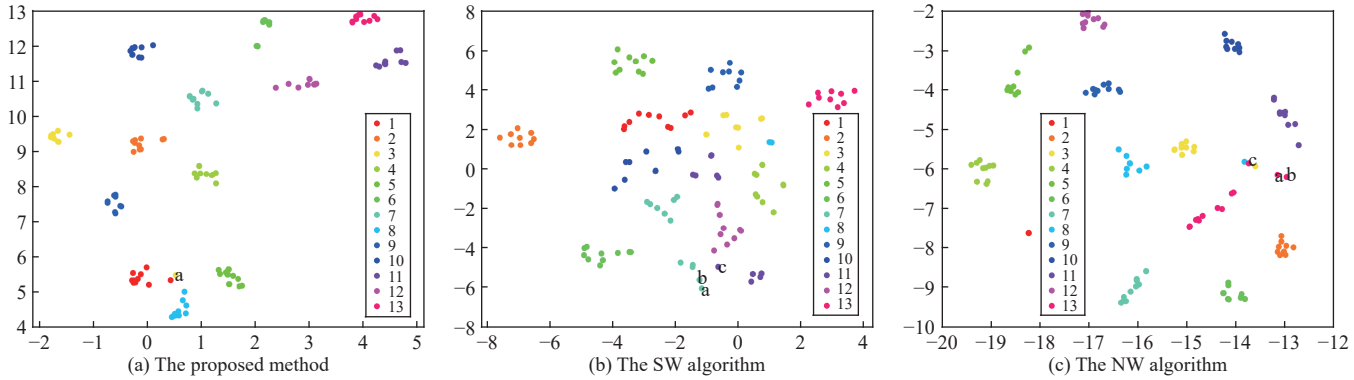

Fig. 2. The clustering results obtained using the proposed method, the SW algorithm, and the NW algorithm.

the other one changing within [1, 10]. It can be observed that the misclassification rate tends to decrease with the increment of either M or c. When M reaches 7, the performance improvement becomes marginal. It is the same with c when it reaches 5. Thus, $M = 7$ and $c = 5$ are used in this case study.

The alarm floods are clustered into 13 groups based on the calculated similarity scores. The clustering results based on the spectral clustering are shown in Fig. 2 (a). According to the original true labels, 129 of the 130 alarm floods are correctly clustered while only 1 is misclassified (marked by "a"). It can be found that the proposed method can effectively identify groups of similar sequences.

Additionally, to demonstrate the advantages of the proposed method, two commonly used sequence alignment algorithms (SW and NW) are exploited for comparison. The clustering results are shown in Figs. 2(b) and 2(c), respectively. According to the original true labels, there were three alarm floods clustered into wrong groups using either the SW or NW algorithm. The misclassified points are marked by "a", "b", and "c". The misclassification rates using the proposed method, SW algorithm, and NW algorithm are shown in Table 1. Comparing with the result using the proposed method, there are more alarm floods misclassified using the SW and NW algorithms. In conclusion, the proposed method is more effective in identifying similar sequences compared to the two classical approaches.

Table 1. Misclassification Rates Using Three Different Methods

	Proposed method	SW	NW
Misclassification rate	0.77 %	2.31 %	2.31 %

Analogous to the SW and NW algorithms, the proposed method has a relative high computational complexity, which is described by $O(mn)$ for pairwise comparison (m and n denote the lengths of two floods). In this case study, the total computational time for 130 alarm floods using the proposed method is 2051 s on a computer with a 2.10 GHz CPU. Therefore, the proposed method might not be suitable in scenarios requiring fast computation, e.g., querying an online sequence in a large database. Even so, the proposed method is still applicable in most cases, especially the offline analysis.

Conclusion: In this letter, a new pattern matching method for alarm flood sequences was proposed based on word embedding and DTW. Through word embedding, alarm messages were encoded into word vectors that represented unique alarms and reflected the relationships between alarm occurrences. Then, groups of similar alarm foods were detected by DTW and spectral clustering. The effectiveness of the proposed method was verified by comparing with classical biological sequence alignment approaches based on the A&E data produced using a public simulation model.

Acknowledgments: This work was supported by the National Natural Science Foundation of China (61903345) and the Knowledge Innovation Program of Wuhan-Shuguang Project (2022010801020208).

References

[1] J. Wang, F. Yang, T. Chen, and S. L. Shah, "An overview of industrial alarm systems: Main causes for alarm overloading, research status, and open problems," *IEEE Trans. Autom. Science and Engineering*, vol. 13, no. 2, pp. 1045–1061, 2016.

[2] *ANSI/ISA-18.2: Management of Alarm Systems for the Process Industries*, Durham, USA: Int. Society of Automation, 2016.

[3] M. R. Parvez, W. Hu, and T. Chen, " Comparison of the Smith-Waterman and Needleman-Wunsch algorithms for online similarity analysis of industrial alarm floods," in *Proc. IEEE Electric Power and Energy Conf.*, 2020, pp. 1–6.

[4] Y. Cheng, I. Izadi, and T. Chen, " Pattern matching of alarm flood sequences by a modified smith-waterman algorithm," *Chemical Engineering Research and Design*, vol. 91, no. 6, pp. 1085–1094, 2013.

[5] W. Hu, J. Wang, and T. Chen, "A local alignment approach to similarity analysis of industrial alarm flood sequences," *Control Engineering Practice*, vol. 55, pp. 13–25, 2016.

[6] M. R. Parvez, W. Hu, and T. Chen, "Real-time pattern matching and ranking for early prediction of industrial alarm floods," *Control Engineering Practice*, vol. 120, p. 105004, 2022.

[7] B. Zhou, W. Hu, K. Brown, and T. Chen, " Generalized pattern matching of industrial alarm flood sequences via word processing and sequence alignment," *IEEE Trans. Industrial Electronics*, vol. 68, no. 10, pp. 10171–10179, 2021.

[8] W. Hu, T. Chen, and S. L. Shah, "Detection of frequent alarm patterns in industrial alarm floods using itemset mining methods," *IEEE Trans. Industrial Electronics*, vol. 65, no. 9, pp. 7290–7300, 2018.

[9] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *CoRR*, pp. 1–9, 2013. DOI: 10.48550/arXiv.1310.4546.

[10] X. Rong, "Word2Vec parameter learning explained," *CoRR*, pp. 1–21, 2014. DOI: 10.48550/arXiv.1411.2738.

[11] A. Lahreche and B. Boucheham, " A fast and accurate similarity measure for long time series classification based on local extrema and dynamic time warping," *Expert Systems with Applications*, vol. 168, p. 114374, 2021.

[12] J. Wen, Y. Xu, and H. Liu, "Incomplete multiview spectral clustering with adaptive graph learning," *IEEE Trans. Cybernetics*, vol. 50, no. 4, pp. 1418–1429, 2020.

[13] G. Yang, W. Hu, W. Cao, and M. Wu, " Simulating industrial alarm systems by extending the public model of a vinyl acetate monomer process," in *Proc. IEEE 39th Chinese Control Conf.*, 2020, pp. 6093–6098.