Research Article

A Data-Driven Approach Based on Historical Hazard Records for Supporting Risk Analysis in Complex Workplaces

Xinhao Wang ^{1,2}, Xifei Huang ³, Yingju Zhang ², Xuhai Pan ¹ and Kai Sheng ^{1,2}

¹Nanjing Tech University, No. 30 Puzhu Road(S), Nanjing 210000, China

²Shenzhen Urban Public Safety and Technology Institute, Fuhua Road, Shenzhen 518000, China

³Shenzhen Green Century Environmental Technology Co. Ltd., 149 Shangbao East Road, Shenzhen 518000, China

Correspondence should be addressed to Kai Sheng; shengkai_risk@hotmail.com

Received 20 July 2021; Accepted 15 October 2021; Published 2 November 2021

Academic Editor: Muhammet Gul

Copyright © 2021 Xinhao Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Risk analysis, as an important prerequisite of risk management, is critical to reducing occupational injuries and other related losses. However, suffering greatly from incomplete hazard identification and inaccurate probability analysis, risk analysis is considered the weakest link in risk management, which seriously affects risk evaluation and control in complex workplaces. To improve the performance of hazard identification and analysis, a data-driven risk analysis approach is established, which consists of an improved equivalent class transformation (Eclat) algorithm, a sliding window model, and a change pattern mining algorithm. Through this approach, a large number of historical hazard records are transformed into association rules composed of object keywords and deviation keywords, and information such as potential keyword combinations, conditional probabilities of potential deviations, and the change pattern of potential hazards can be extracted. The function of the approach is threefold. Firstly, the data-driven risk analysis process is designed to identify the association rules between different hazard keywords. Secondly, Eclat algorithm is optimized to calculate the frequency and probability of potential hazards, which is conducive to improving the accuracy of probability estimation. Thirdly, the change pattern is developed to analyse the hazard change trend to support the cause analysis. A practical application in a Chinese hazardous chemical manufacturer is presented. Case studies have shown that the efficiency of the improved algorithm is increased by 13.68%, and 59.66% of potential hazards can be identified in advance, and relevant information can be extracted to support risk analysis.

1. Introduction

Preventing and mitigating accidents, protecting employees from occupational injuries, and protecting the environment from damage is a major goal of all industries and superintendents [1]. Risk analysis, as an important prerequisite of risk management, is critical to reducing injuries and other related losses [2]. However, suffering greatly from incomplete hazard identification and inaccurate probability analysis, risk analysis is considered the weakest link in the risk management process, which seriously affects risk evaluation and control in complex workplaces [3, 4]. Improving the ability of superintendents to identify and analyse hazards in complex workplaces will be necessary to address the issues raised by ongoing accident occurrences [5].

To achieve this goal, a large number of methods are designed and used, which can be divided into administrative and technological [6]. Administrative methods mainly rely on the experience of superintendents and historical records, such as hazard and operability analysis [7], failure mode and effects analysis [8, 9], layer of protection analysis [10, 11], job hazard analysis [12], safety checklists [13], and fault tree [14]. In recent years, many researchers have gradually realized that the results of the administrative methods are not accurate enough due to the corresponding hazards that are analysed based on information with fuzziness and randomness [15]. Meanwhile, with the complexity of the workplace, it will be difficult for superintendents to identify hazards completely. The workplace has reached saturation with respect to traditional safety strategies that were originally implemented to comply with regulations [12].

In recent years, technological methods represented by data mining have been aided in solving some specific problems in various fields of safety engineering. The articles can be divided into two different types: articles focusing on textual data and articles focusing on numerical data [16]. Textual data mainly consist of accident and incident records. Tixier analysed textual data of construction accidents and identified combinations of accident causes by utilizing natural language processing (NLP) [17]. Brown and Yang used latent Dirichlet allocation (LDA) and text-based Bayesian network (TBN) to analyse railway safety text data to support typical accident risk analysis [18, 19]. Marucci-Wellman studied the classification method of narrative text based on machine learning, which is used to classify a large number of occupational injury and illness events [20]. Mutlu and Altuntas integrated FMEA, FTA, and BIFPET methods to analyse occupational safety and health risks [21–23]. Robinson applied the latent semantic analysis method to analyse the main causal factors of the safety narrative [24].

Due to the lack of standard terms and formats, there are few algorithms that can analyse textual data directly. For this reason, in some articles, textual data is transformed into keywords combinations or numerical data to adapt to more numerical data mining algorithms. Lu extracted reciprocating compressor information containing failure symptoms and causes of failed components from technical journal articles and analysed the association rules between monitoring parameters and maintenance records [25]. Xu used a priori algorithm to analyse the main factors leading to road traffic accidents and their association relationships [26]. Song applied a Bayesian network to explicitly explore statistical associations between crash severity and significant variables [27]. In addition, grey comprehensive correlation degree, artificial neural network (ANN), random forests (RF), K-nearest neighbour (KNN), hidden Markov model (HMM), and other algorithms are also used to analyse spare parts [28], turn-back faults in urban rail [29], cause of traffic accident [30], human unsafe factors [31], and so on.

The above-mentioned articles show that converting textual data into several keywords and applying association rule algorithms for further analysis are excellent ideas for obtaining risk information. However, there are still some small weaknesses in these articles, which cannot be directly applied to workplace hazard identification. Firstly, these existing articles are mainly applied to specific scenarios or specific types of accidents, such as reciprocating compressor failures, road traffic accidents, and so on, which have a common feature that the contribution factors are less and fixed. There are many managed objects in complex workplaces, including equipment, facilities, personnel, processes, and so on, leading to complex and diverse hazards. It will be time-consuming to analyse and identify all types of hazards one by one using the above-mentioned methods. Therefore, the research object should be extended to multiple types of hazards, and the information of potential hazards should be obtained as much as possible through the association relationship. Secondly, the importance and frequency of hazards are quite different, which is often ignored in existing studies. Traditional association rule algorithms tend to

ignore hazards with high consequences and low frequency, so an improved algorithm needs to be developed. Thirdly, workplace hazards are dynamically changing, which is different from the contribution factors of specific accidents in the above-mentioned articles. It is necessary to design an algorithm for analysing hazard trends.

The aim of this article is to apply data mining algorithms to improve the ability of superintendents to identify and analyse hazards in complex workplaces. In this article, hazard identification and analysis can be seen as building and analysing a complex network. Hazard records can be simplified as the combination of several keywords. For example, “the pipeline is corroded” can be decomposed into “pipeline” and “corroded.” If “pipeline” and “corroded” are represented by two nodes, “the pipeline is corroded” can be represented by two nodes and their connecting lines. Several nodes and their connecting lines are combined to form a complex network that can reflect the hazard information of the workplace. The purpose of hazard identification can be regarded as finding potential nodes through known nodes. Hazard analysis is to calculate the probability of potential nodes and predict the change trend of these nodes. Motivated by these ideas, a data-driven risk analysis approach is designed, which consists of an improved equivalent class transformation (Eclat) algorithm, sliding window model, and change pattern mining algorithm, to discover the nodes, connections, and their changes.

The function of this approach is threefold. Firstly, the data-driven risk analysis process is designed to identify the association rules between different hazard keywords and solve the problem of incomplete hazard identification caused by insufficient experience. Secondly, Eclat algorithm is optimized to calculate the frequency and probability of potential hazards, which is conducive to improving the accuracy of probability estimation. Thirdly, the change pattern is developed to analyse the hazard change trend to support the cause analysis.

The article is structured as follows. Section 2 presents the research background about Eclat algorithm and sliding window model. Section 3 presents the workflow and main links of the data-driven risk analysis approach. Section 4 is a case study about a Chinese hazardous chemical manufacturer, which is used to analyse the practicality of the approach. Conclusions and future work are given in Section 5.

2. Related Work

2.1. Equivalent Class Transformation Algorithm. Association rules are proposed to mine the relationship between elements in a large database, which are usually expressed in the form of $\{A \rightarrow B\}$ [32]. Association rule mining is divided into two stages. Firstly, all frequent items meeting the minimum support threshold are found. And then the rules meeting the minimum confidence threshold are identified from the frequent item sets. Support is the co-occurrence probability of elements, that is, the probability at which the element A and the element B appear together (equation (1)). Confidence is the conditional probability, that is, the probability that the element B exists if the element

A exists (equation (2)). In this article, the association rules mining algorithm is used to analyse the association relationship between different hazard keywords and calculate the probability of keywords to support risk analysis.

$$\text{Support}(A, B) = P(A \cup B) = \frac{\text{Count}(A \cup B)}{\text{Count}(\text{all})}, \quad (1)$$

$$\text{Confidence}(A \longrightarrow B) = \frac{\text{Support}(A, B)}{\text{Support}(A)}. \quad (2)$$

Eclat algorithm is used to mine the association relationship in hazard records. Eclat algorithm is an algorithm that uses vertical data format to search frequent items [33]. Taking hazard identification as an example, seven hazards were identified, and the corresponding keywords were extracted. The data formats of Tables 1 and 2 are called horizontal data format and vertical data format, respectively.

Based on the vertical data format, the hazard sets are divided into subsets by using the equivalence relationship of keyword ID. The bottom-up search method is used in each subset to generate frequent items independently. Figure 1 shows the search path when the minimum support threshold is 3. Since there is no need to calculate and scan the candidate frequent items, Eclat is suitable for analysing hazard datasets, which include many types and quantities.

Furthermore, in complex workplaces, the number and frequency of keywords are quite different. Some keywords, especially those related to hazards with high severity and low frequency, will be missed by the traditional algorithm. To overcome this defect, the concept of multiple minimum support thresholds needs to be introduced, which can improve the number of association rules by reducing the minimum support threshold of specific keywords.

2.2. Sliding Window Model. In safety engineering, the objects of data mining can be divided into static data and stream data. Hazard identification data can be classified as stream data. Stream data is a large, continuous, fast, and time-varying data sequence. Let t be the time stamp and A_t be the data collected at t , then the stream data can be expressed as $\{\dots, A_{t-1}, A_t, A_{t+1}, \dots\}$.

Stream data is unbounded, in which only limited data can be processed, mined, and analysed. Thus, the definition of the data stream window is introduced [34]. The sliding window model is a typical model for data stream processing. Its core idea is that when processing the data stream, users often care about the recent data. Let the window size be T , which can be one second, one minute, one day, and the time point be t , then the range of the window is $[t - T, t]$. When the time point t changes in sequence, the window will slide, as shown in Figure 2.

Based on the sliding window model, the historical hazard records can be analysed to obtain the hazard information of the current workplace. Through the comparison of hazard records in different stages, the change of hazards will also be pointed out.

TABLE 1: Horizontal data format.

Hazard ID	Keyword set
Hazard 1	{A, B, C, D}
Hazard 2	{B, D}
Hazard 3	{A, B}
Hazard 4	{C, D}
Hazard 5	{A, B, C, D}
Hazard 6	{B, D}
Hazard 7	{B, C, D}

3. Methodology

The data-driven risk analysis approach can be divided into three parts, association rule mining, change pattern mining, risk analysis and confirmation, as shown in Figure 3. Firstly, NLP is applied to transform textual hazard records into a combination of several hazard keywords, and the improved Eclat algorithm is used to mine the association rules between these keywords. Then the association rules in different stages are divided into two groups, and the hazard change patterns, such as increase pattern, decrease pattern, and new pattern, are identified. Finally, the target object is mapped to the association rule set and the change pattern set; the potential hazards and their conditional probabilities are extracted; and the change trend of the potential hazards is also informed. In practice, the data-driven risk analysis approach can be completed by an information system. As long as the object or scope is clear, the corresponding risk information will be automatically informed.

3.1. Association Rule Mining

3.1.1. Data Transformation. Data transformation is the process of transforming textual hazard records into structured data. In this process, each hazard record will be transformed into several keywords, and each keyword will be assigned an index and frequency. In this way, the textual hazard records can be quantitatively analysed.

There are various types of hazards, and the ways of hazard description are also different. In China, the interim provisions on the investigation and control of safety accidents require enterprises to carry out hazard identification, focusing on the dangerous state of equipment and facilities, human unsafe behaviours, and management defects. Standardization of enterprise safety production (GB/T 33000-2016) provides enterprises with a standardized list to guide enterprises to carry out this work. Its core idea is to identify and eliminate the deviation between equipment, facilities, personnel behaviours, management, other objects, and the standard state. Figure 4 shows the description of the four types of hazards. By analysing the content of the text, it can be found that the keywords of different hazards can be divided into entities, attributes, and deviations. Entities include equipment, facilities, workplaces, and so on. Attributes include process parameters, hazard factors, and so on. These two kinds of keywords are the objects of hazard identification and risk analysis. Deviation refers to the

TABLE 2: Vertical data format.

Keyword ID	Hazard set
A	{Hazard 1, Hazard 3, Hazard 5}
B	{Hazard 1, Hazard 2, Hazard 3, Hazard 5, Hazard 6, Hazard 7}
C	{Hazard 1, Hazard 4, Hazard 5, Hazard 7}
D	{Hazard 1, Hazard 2, Hazard 4, Hazard 5, Hazard 6, Hazard 7}

FIGURE 1: Search process of Eclat algorithm.

FIGURE 2: Sliding window model.

deviation of the object from the safety standard, such as defect, failure, low, exceed, and so on.

Therefore, a hazard record can be transformed into one or more object keywords and one deviation keyword. If there are multiple deviation keywords in a hazard record, it can be divided into multiple combinations. To improve the accuracy of hazard analysis, these keywords need to be standardized. For example, “grounding wire” and “earth wire” are transformed to “ground lead”; “exceed the standard” and “higher than the threshold” are transformed to “more.” The expression is as follows:

$$\text{Hazard} = \{\text{Object 1, Object 2, } \dots, \text{Deviation 1}\}. \quad (3)$$

Several hazards are sorted according to the identified time, and the hazard stream data is formed. According to the sliding window model, the stream data are divided into two

datasets for further analysis [35]. As is shown in Table 3, stage 1 is closer to the current period, and the operating conditions and hazards in this stage are also similar to the actual situation. Therefore, the hazards identified in stage 1 can be used to analyse the potential hazards that may exist in the current stage. By comparing the difference between the hazards identified in stages 1 and 2, the trend of hazard changes can be discovered.

THU Lexical Analyzer for Chinese (THULAC) is used to decompose hazard records, and a self-made Safety Engineering Terminology Dictionary (SETD) is used to extract and standardize keywords. All these works will be done automatically by Python.

3.1.2. Improved Eclat Algorithm. Eclat algorithm is applied to mine the association relationship between the object

FIGURE 3: Process of data-driven risk analysis approach.

- Record 1: The safety administrator found that the pulley of some metal cutting machine tools was not installed with protective cover, which should be rectified as soon as possible.
 entity entity deviation entity
- Record 2: The monitoring platform shows that the liquid level of the underground slop tank in the diesel oil hydrofining plant is low.
 attribute entity entity deviation
- Record 3: The monitoring report shows that the electromagnetic radiation in the non-destructive workshop exceeds the standard.
 attribute entity deviation
- Record 4: Check the oil depot and find some cracks on the road surface.
 entity deviation entity

FIGURE 4: Keywords in typical hazard records.

TABLE 3: Example of sliding window model.

Stage	Time horizon	Hazard dataset
Stage 1	[2021-05-01, 2021-07-01)	{Hazard 1, Hazard 2, Hazard 3, Hazard 4, Hazard 5, Hazard 2, Hazard 3, Hazard 4}
Stage 2	[2021-03-01, 2021-05-01)	{Hazard 8, Hazard 7, Hazard 2, Hazard 6, Hazard 3, Hazard 4, Hazard 5}

keywords and the deviation keywords, and the following expression can be obtained:

$$\text{Hazard } X = \{\text{Object } Y, \text{Deviation } Z\}, \quad (4)$$

$$\text{Support}(\{\text{Object } Y, \text{Deviation } Z\}) = \frac{\text{Count}(\text{Object } Y \cup \text{Deviation } Z)}{\text{Count}(\text{all Hazard})}, \quad (5)$$

$$\text{Confidence}(\{\text{Object } Y \longrightarrow \text{Deviation } Z\}) = \frac{\text{Support}(\{\text{Object } Y, \text{Deviation } Z\})}{\text{Support}(\{\text{Object } Y\})}. \quad (6)$$

Equation (5) reflects the probability of Hazard X in the workplace. Through this equation, the frequency of occurrence of Hazard X can be calculated, which can be used to assist the superintendent in determining the key objects. Equation (6) reflects the probability of Deviation Z among all hazards related to Object Y . Different methods and tools are needed to identify different deviations. Equation (6) can be used to assist superintendents in formulating identification strategies in advance.

However, in the same workplace, the frequency and severity of different hazards are different. The higher the severity of the hazards, the lower the frequency, which leads to lower support for the corresponding keywords. In association rule mining, if the minimum support threshold is set too high, some important keywords with lower support will not be extracted. Conversely, if the minimum support

threshold is set too low, the frequent keyword sets will increase dramatically, which reduces the density of hazard information and increases the difficulty of hazard analysis.

In response to this phenomenon, the concept of multiple minimum support thresholds is proposed to improve the Eclat algorithm. The definition is as follows.

Definition 1. Let the keyword set of all hazard records be represented as $I = \{\text{Object } 1, \dots, \text{Deviation } 1, \dots\}$ and use $\min_sup(\{\text{Object } x\})$ and $\min_sup(\{\text{Deviation } x\})$ to represent their minimum support thresholds. Then for $X = \{\text{Object } x, \dots, \text{Deviation } x, \dots\}$, when $X \subseteq I$ and $|X| \geq 2$, the minimum support threshold of X is equal to the minimum value of the minimum support thresholds of all keywords. The expression is as follows:

$$\min_sup(X) = \min\{\min_sup(\{\text{Object } x\}), \dots, \min_sup(\{\text{Deviation } x\}), \dots\}. \quad (7)$$

Definition 2. Let $\min_conf$ be the minimum confidence threshold. The conditions of the association rule $\{\text{Object } y \rightarrow \text{Deviation } z\}$ can be expressed as follows:

$$\begin{aligned} \text{Support}(\{\text{Object } y\}) &\geq \min_sup(\{\text{Object } y\}), \\ \text{Support}(\{\text{Object } y, \text{Deviation } z\}) &\geq \min_sup(\{\text{Object } y, \text{Deviation } z\}), \\ \text{Confidence}(\{\text{Object } y \rightarrow \text{Deviation } z\}) &\geq \min_conf. \end{aligned} \quad (8)$$

After setting the multiple minimum support thresholds, the recursive nature of the Eclat algorithm is destroyed, which means that the subset of frequent items may not meet their minimum support thresholds. Take Tables 1 and 2 as an example, let the minimum support thresholds of A, B, C, and D be 2, 6, 5, and 6 respectively, then $\{A\}$, $\{B\}$, $\{D\}$, $\{A, B\}$, and $\{A, D\}$ can be identified by the Eclat algorithm, as shown in Figure 5(a). Although $\{A, C\}$, $\{A, B, C\}$, $\{A, B, D\}$, $\{A, C, D\}$, and $\{A, B, C, D\}$ also meet their minimum support thresholds, they are still omitted by Eclat algorithm, which is mainly because their subsets, such as $\{C\}$, $\{B, C\}$, $\{B, D\}$, $\{C, D\}$, and $\{B, C, D\}$, do not meet their minimum support thresholds.

For the above situation, sort 1-items in ascending order according to their minimum support thresholds. Then the n -items ($n \geq 2$) with the same prefix have the same minimum support threshold, which is equal to the minimum support threshold of the corresponding prefix. As shown in Figure 5(b), $\min_sup(\{A\}) = \min_sup(\{A, C\}) = \min_sup(\{A, B, C\}) = \min_sup(\{A, B, C, D\}) = 2$. In each of these branches, the subset of frequent n -items ($n \geq 2$) are also frequent, so the recursive property will be restored.

The search process is as follows:

Step 1: traverse the hazard dataset and construct the vertical data format, as shown in Table 2

Step 2: according to the multiple minimum support thresholds, sort the keywords in ascending order and construct the data space as shown in Figure 5(b)

Step 3: search each of these branches to obtain keyword sets that meet their minimum support thresholds

The search process can be summarized as: for any keyword set $X \subset Y$, when X and Y have the same prefix, if $\text{Support}(X) \geq \min_sup(X)$, then $\text{Support}(Y) \geq \min_sup(Y)$. The proof is as follows:

Set $X = \{x_1, x_2, \dots, x_n\}$, $Y = \{y_1, y_2, \dots, y_n\}$, then $X \subset Y$ can deduce $n > m$.

Because X and Y have the same prefix and 1-items are arranged in ascending order of their minimum support thresholds, $\min_sup(X) = \min_sup(\{x_1\}) = \min_sup(\{y_1\}) = \min_sup(Y)$, which means the keyword sets with the same prefix have the same minimum support threshold.

It can be inferred from $X \subset Y$ that $\text{Support}(X) \geq \text{Support}(Y)$. Then, when X does not meet the minimum support threshold, $\min_sup(X) = \min_sup(Y) > \text{Support}(X) \geq \text{Support}(Y)$. In other words, Y does not meet the minimum support threshold.

3.2. Change Pattern Mining. Change pattern mining is a method to analyse the change trend of the research object by

FIGURE 5: Comparison of the search process. (a) Search process of Eclat. (b) Search process of improved Eclat.

comparing the differences of the association rules in two stages [36]. This method has been successfully applied in supermarket sales to analyse changes in customer consumption patterns [37]. At present, the defined change patterns mainly include added patterns, subtractive patterns, emerging patterns, and unexpected patterns, which reflect different changes. However, the change patterns of hazard and consumption are not consistent. The change patterns need to be redefined.

The hazard change patterns can be divided into growth pattern, decrease pattern, new deviation pattern, new object pattern, and new hazard pattern. Related terms and symbols are defined as follows:

R_i^{W1} : $\{O_i^{W1} \longrightarrow D_i^{W1}\}$ is an association rule in W_1 ; O_i^{W1} ($|O_i^{W1}| \geq 1$) and D_i^{W1} ($|D_i^{W1}| = 1$) are object and deviation in W_1 ; and the support of R_i^{W1} is $\text{Support}(R_i^{W1})$

R_j^{W2} : $\{O_j^{W2} \longrightarrow D_j^{W2}\}$ is an association rule in W_2 ; O_j^{W2} ($|O_j^{W2}| \geq 1$) and D_j^{W2} ($|D_j^{W2}| = 1$) are object and deviation in W_2 ; and the support of R_j^{W2} is $\text{Support}(R_j^{W2})$

Left_{ij} is the similarity of O_i^{W1} and O_j^{W2} , where $\text{Left}_{ij} = (|O_i^{W1} \cap O_j^{W2}| / |O_i^{W1}|)$

Right_{ij} is the similarity of D_i^{W1} and D_j^{W2} , where $\text{Right}_{ij} = \begin{cases} 1 & (D_i^{W1} = D_j^{W2}) \\ 0 & (D_i^{W1} \neq D_j^{W2}) \end{cases}$

Rule_{ij} is the similarity of R_i^{W1} and R_j^{W2} , where $\text{Rule}_{ij} = \begin{cases} \text{Left}_{ij} & (D_i^{W1} = D_j^{W2}) \\ 0 & (D_i^{W1} \neq D_j^{W2}) \end{cases}$

Left_i is the maximum of Left_i , where $\text{Left}_i = \{\text{Left}_{i1}, \text{Left}_{i2}, \dots, \text{Left}_{in}\}$

Right_i is the maximum of Right_i , where $\text{Right}_i = \{\text{Right}_{i1}, \text{Right}_{i2}, \dots, \text{Right}_{in}\}$

Rule_i is the maximum of Rule_i , where $\text{Rule}_i = \{\text{Rule}_{i1}, \text{Rule}_{i2}, \dots, \text{Rule}_{in}\}$

ST is the similarity threshold

Let W_1 be close to the current time point, the criteria for hazard change patterns are shown in Table 4.

Growth pattern: this pattern states that an association rule appears in both W_1 and W_2 , and compared to W_2 , the support of this rule in W_1 has increased. For the hazard change patterns, if the support of an association rule remains unchanged, it is also regarded as a growth pattern. This pattern usually reflects that the hazard has not been effectively controlled, resulting in the same hazard being continuously identified in W_1 . In this case, the main task of the superintendent is to formulate and implement a risk management plan.

Decrease pattern: this pattern is the opposite of the growth pattern. Compared to W_2 , the support of the rule is reduced in W_1 . Superintendents should continue to implement the risk management plan.

New deviation pattern: compared to W_2 , some new deviations are identified in the same object. For example, in W_2 , the main deviation of the pipeline was corrosion, but in W_1 , there was a leak in the same pipeline. The lack of safety management on specific objects is the main reason for the emergence of new deviations. In this case, it is urgent for superintendents to strengthen safety management for the object.

New object pattern: compared to W_2 , some new objects are identified with similar deviations. For example,

TABLE 4: The criteria for hazard change patterns.

Change pattern	Left _i ≥ ST	Right _i = 1	Rule _i ≥ ST	Support ($R_i^{W_1}$) ≥ Support ($R_j^{W_2}$)
Growth pattern	Yes	Yes	Yes	Yes
Decrease pattern	Yes	Yes	Yes	No
New deviation pattern	Yes	Yes/no	No	—
New object pattern	No	Yes	—	—
New hazard pattern	No	No	—	—

“corroded pipeline” was identified in W_2 , and “corroded electrostatic jumper” was also identified in W_1 . Poor workplace environment or process problems are generally the cause of similar deviations in various objects. Environmental management and process quality control are the keys to eliminating this phenomenon.

New hazard pattern: for the hazard in W_1 , the object and deviation are not identified in W_2 . The main reason is the change of equipment, facilities, personnel, or process. PHA, HAZOP, and other administrative methods can be used to analyse the risk to prevent the deterioration of these hazards.

3.3. Risk Analysis and Confirmation. Through the above two links, hazard association rules and hazard change patterns are extracted. In the process of daily hazard identification, the superintendent can determine the objects, such as pipelines, storage tank, or power supply systems, and map them into the association rule set and the change pattern set. Then, the association rules related to the object can be extracted, and the possible deviation and change trend will be obtained to improve the efficiency of risk analysis.

4. Application and Analysis

4.1. Data. The hazard data comes from the workplace hazard identification records of a Chinese hazardous chemical manufacturer, as shown in Table 5. General Code for Safety Standardization of Hazardous Chemicals Manufacturers is used as the work guideline of hazard identification, which includes basic management, equipment and facilities, technological process, working environment, and occupational hazards. Basic management includes work permit, change management, contractor and supplier management, emergency management, and so on. Equipment and facilities mainly include production facilities, safety facilities, and related equipment. The technological process is different in different enterprises, and the main parameters include temperature, pressure, flow, liquid level, composition, and so on. The working environment includes workshop environment, factory environment, warehouse environment, and so on. Occupational health mainly includes occupational hazard management, occupational disease control, and occupational health monitoring.

From 2017 to 2020, 152 hazard identification activities were carried out, and a total of 12,353 hazards were identified and recorded in the form of Table 5. The types of hazards are shown in Figure 6. These 152 sets of hazard data will be used in this article.

4.2. Application

4.2.1. Hazard Association Rule Mining. Ten hazard datasets (numbered 143 to 152) in the third quarter of 2020 are selected as the observation group for simulation application. Forty datasets (numbered 103 to 142) in the first half of 2020 and the second half of 2019 are divided into W_1 and W_2 , respectively, which slide with the change of the observation time, as shown in Figure 7. In W_1 , association rules are mined to identify the potential hazards and calculate the conditional probability. In W_2 , the association rules are compared with those in W_1 to analyse the hazard change patterns. Experimental data are used to calculate the accuracy of hazard prediction.

Python 3.7 is used to realize the data-driven risk analysis approach. For historical hazard association rule mining, the multiple minimum support thresholds of each keyword are set to 20% of its frequency, and the minimum confidence threshold is 0.2. Around 684 historical hazard association rules in W_1 and 784 historical hazard association rules in W_2 are obtained.

The left items (i.e., objects) are clustered in the association rules, and the association rules are extracted with higher frequency and confidence, as shown in Table 6. The first item indicates that the deviation of the “warming apparatus” and the other two objects in W_1 is “missing,” the frequency is 10, and the confidence is 0.67. In W_2 , the main deviation types of the “warming apparatus” and the other two objects are “eroded,” “missing,” and “damaged.” Through Table 6, it can be found that the frequency and confidence of the deviations corresponding to “warming apparatus,” “safety cages,” “rain sewage valve,” “stop valve,” “discharge outlet,” and “environmental monitor” are higher, and these should be checked first.

The visualization of the main association rules is shown in Figure 8. It can be seen from Figure 9 that the main deviations in W_1 are “missing,” “more,” and “inefficient,” and the main deviations in W_2 are “missing,” “damaged,” and “more.” Compared with Figure 8(b), the type of nodes and the number of lines in Figure 8(a) are different, which reflects that the hazard has changed in W_1 .

4.2.2. Hazard Change Pattern Mining. The association rules in W_1 and W_2 are compared in Figure 9. In Figure 9, the abscissa and the ordinate are used to represent the object and deviation, respectively. The intersection of the object and the deviation is regarded as a type of hazard, where the hazards in W_1 are represented by the red circles, and the hazards in W_2 are represented by the blue circles. The frequency of

TABLE 5: Hazard record for the Chinese hazardous chemical manufacturer.

Type	Hazard information
Time	2020-05-11
Location	Transformer room
Hazard type	Equipment and facilities
Hazard description	It was found that no safety warning signs were posted around the transformer
Responsible person	Team leader of powerhouse

FIGURE 6: The types of hazards.

FIGURE 7: Hazard window partition.

TABLE 6: The main association rules in W_1 and W_2 .

NO.	Main objects	Main deviations in W_1 and W_2
1	Warming apparatus + 2 objects	Missing (10, 0.67, W_1); eroded (9, 0.60, W_2); missing (8, 0.53, W_2); damaged (6, 0.40, W_2)
2	Safety cages + 9 objects	Missing (7, 1.00, W_1); missing (10, 1.00, W_2); more (7, 0.53, W_2); damaged (2, 0.25, W_2)
3	Rain sewage valve + 6 objects	Missing (6, 1.00, W_1); inefficient (2, 1.00, W_1); inefficient (3, 0.75, W_2)
4	Stop valve + 5 objects	Missing (6, 0.86, W_1); inefficient (2, 0.67, W_1); substandard (2, 0.50, W_1)
5	Discharge outlet + 6 objects	Missing (6, 0.75, W_1); more (6, 0.67, W_1); inefficient (4, 0.57, W_1)
6	Environmental monitor + 7 objects	Missing (6, 0.86, W_1); missing (6, 0.50, W_2); more (2, 0.50, W_2)
7	Control index + 5 objects	More (5, 0.63, W_1); inefficient (2, 1.00, W_1); discrepant (2, 0.29, W_2)
8	Drainage + 3 objects	Missing (5, 0.63, W_1); polluted (2, 0.50, W_1); inefficient (2, 0.50, W_1); leaked (2, 0.50, W_1)
9	Diversion facilities + 5 objects	Missing (5, 0.71, W_1)
10	Sulfur dioxide + 7 objects	More (4, 0.67, W_1); substandard (2, 0.67, W_1); inefficient (2, 0.50, W_1); inefficient (2, 0.67, W_2); more (2, 0.67, W_2)
11	Monitoring data + 6 objects	Inefficient (4, 0.57, W_1); more (2, 1.00, W_1); more (2, 0.67, W_2)
12	Steady rate + 10 objects	More (4, 0.80, W_1); missing (7, 0.70, W_2); more (6, 1.00, W_2)
13	Installation height + 5 objects	Less (3, 0.75, W_1); substandard (2, 1.00, W_2); less (2, 0.33, W_2)
14	Interlock circuit + 5 objects	Substandard (2, 1.00, W_1); discrepant (2, 0.67, W_1); substandard (2, 1.00, W_2)
15	Boiler + 5 objects	Leaked (2, 1.00, W_1); more (2, 1.00, W_1); less (2, 0.67, W_1)
16	Lead sealing + 8 objects	Closed (2, 0.50, W_1)
17	Oil trap + 6 objects	More (2, 1.00, W_1); missing (8, 1.00, W_2); more (2, 1.00, W_2)
18	Chemical oxygen demand + 7 objects	More (2, 1.00, W_1); more (5, 0.83, W_2); broken-down (2, 0.50, W_2)
19	Threading apparatus + 4 objects	Damaged (2, 1.00, W_1); more (3, 1.00, W_2); damaged (2, 1.00, W_2)
20	Equipotential connection + 8 objects	Missing (18, 0.75, W_2); unmatched (6, 0.55, W_2); eroded (6, 0.55, W_2)

TABLE 6: Continued.

NO.	Main objects	Main deviations in W_1 and W_2
21	Lead sealing + 8 objects	Missing (9, 0.90, W_2); damaged (7, 0.70, W_2)
22	Sewage valve + 6 objects	Missing (8, 1.00, W_2); damaged (2, 0.50, W_2); more (2, 0.33, W_2)
23	Equipment enclosure + 8 objects	Missing (8, 0.73, W_2); unmatched (3, 1.00, W_2); damaged (3, 1.00, W_2)
24	Diversion facilities + 5 objects	Missing (7, 0.78, W_2); leaked (2, 0.50, W_2); less (2, 0.50, W_2)
25	Flame detector + 4 objects	Missing (6, 1.00, W_2); unmatched (2, 1.00, W_2); less (2, 1.00, W_2); more (2, 1.00, W_2); substandard (2, 1.00, W_2)
26	Fire pump + 4 objects	Missing (6, 0.67, W_2); inefficient (3, 0.43, W_2); broken-down (2, 1.00, W_2)
27	Installation site + 3 objects	Leaked (2, 1.00, W_2); less (2, 1.00, W_2); substandard (2, 1.00, W_2)
28	Cement + 3 objects	Eroded (2, 1.00, W_2); damaged (2, 1.00, W_2); inefficient (2, 0.67, W_2)

FIGURE 8: Historical hazard association rules visualization. (a) Rule in W_1 . (b) Rule in W_2 . FIGURE 9: Comparison of association rules in W_1 and W_2 .

hazards is reflected by the magnitude, and the probability of occurrence of deviation is reflected by transparency. The higher the frequency and probability, the larger the corresponding shape and the darker the color. As shown in Figure 9, although the main types of deviations in the two stages are similar, the objects of the hazards are quite different.

To quantify the difference of hazards and identify the hazard change trend, hazard change pattern mining needs to be carried out. Let ST be 0.7, and the results of hazard change pattern mining are shown in Table 7 and Figure 10.

New object patterns and new deviation patterns accounted for 93%, which reflects the inadequate management of specific objects and the workplace environment. For example, due to the inadequate management of the drainage system, the safety signs of rain sewage valves and discharge outlets are lost, and the drainage efficiency is insufficient. This situation is also reflected in Figure 9, that is, the red circle corresponding to the rain sewage valve and the discharge outlet does not overlap the blue circle.

Less growth patterns and decrease patterns reflect that superintendents have a stronger ability in hazard control. Once a hazard is identified, the most similar hazards can be quickly eliminated. However, the number of hazards such as the “warming apparatus, missing” and “sulfur dioxide, more” is on the rise. In Figure 9, they are represented as larger red circles that overlap the blue circles. It is necessary for superintendents to carry out cause analysis and optimize control measures.

The reason there is no new pattern is that the hazardous chemical manufacturer has neither introduced new equipment nor changed the process in the past two years.

4.2.3. Simulate the Risk Analysis Process. In risk analysis, superintendents will obtain relevant risk information as long as the identified object is mapped into the association rule set and change pattern set. For example, in activity 143, the “hydraulic supply line” was mapped into the dataset, and the following results were output, as shown in Table 8 and Figure 11.

The number of “warming apparatus, missing” is the highest, and the probability is 71%, which is a large number. It is recommended that the warming apparatus should be checked first. Moreover, this hazard is a growth pattern, which means that the frequency is increasing. Therefore, even if this type of hazard is not identified in this activity, it is necessary to take precautionary measures.

The frequency and probability of “flange, bolt, and missing” are also high, and it is a new object pattern, indicating that the maintenance of flanges is insufficient. Therefore, it is recommended to carry out centralized maintenance of flanges and pay attention to the safety status of bolts.

“Bracket, unreliable” and “valve, eroded” are the new deviation patterns, which may be related to weather factors. In the first half of 2020, rainstorms and gales occurred many

times in this area. And it is predicted that this weather will continue until October, which will lead to a further increase in the frequency of these hazards. Superintendents should pay attention to the impact of weather on equipment and facilities.

Finally, a total of 54 hazards were predicted in this activity.

4.3. Analysis

4.3.1. Efficiency Analysis. Apply multiple minimum support thresholds and a single minimum support threshold to mine frequent item sets. And then, calculate the number of keywords in frequent item sets, the number of frequent items, and the ratio between them. The ratio of keywords to frequent items is used to express the value density of frequent item sets. The higher the ratio, the more hazard information contained in frequent item sets, which is beneficial to improve the efficiency of association rule analysis. The hazard datasets in W_1 are used for analysis, where the multiple minimum support thresholds are set according to 20% of the frequency of each keyword, and the single minimum support threshold is set to 2, 5, and 10. The results are shown in Table 9 and Figure 12.

In Figure 12, the number and ratio of the keywords are represented by the size and position of the circle. When the single minimum support threshold is equal to 5 and 10, the number of keywords has significantly decreased, which is not enough to support risk analysis. When the single minimum support threshold is equal to 2, the number of keywords is the same as the first group, but the ratio is about 10% lower than the first group on average, and the total ratio is 13.68% lower. Therefore, it can be considered that setting multiple minimum support thresholds has a better effect on improving the efficiency of association rule mining.

4.3.2. Application Effect. Analyse the ten experimental datasets from 143 to 152, count the number of predicted hazards, and calculate the ratio of the predicted number to the actual number, as shown in Figure 13. In the experimental analysis, an average of 38.6 hazards were predicted, accounting for 59.66%. This shows that in each hazard identification activity, 59.66% of the potential hazards can be notified to the superintendent in advance, which not only improves the pertinence of hazard identification but also provides basic information for risk evaluation and control.

The simulation application shows that the data-driven risk analysis method can predict more than half of the hidden dangers so as to improve the hazard identification efficiency in the workplace. However, there are still some limitations in practical application. Firstly, the method is highly dependent on historical data. Insufficient historical data and inaccurate data records will reduce the prediction accuracy. Secondly, this method can only predict and analyse the identified hazards based on historical records, and there is no analysis means for the unrecorded hazards.

TABLE 7: Results of hazard change pattern mining.

Change pattern	Number	Ratio (%)
New deviation pattern	205	30
New object pattern	430	63
Decrease pattern	17	5
Growth pattern	32	2

FIGURE 10: Proportion of hazard change pattern.

TABLE 8: The output of the data-driven risk analysis approach.

Input	No.	Potential object	Potential deviation	Frequency	Confidence	Change pattern
Hydraulic supply line	1	Warming apparatus	Missing	10	0.71	Growth pattern
	2	Flange and bolt	Missing	10	0.67	New object pattern
	3	Bracket	Unreliable	2	0.29	New deviation pattern
	4	Valve	Eroded	2	1.00	New deviation pattern

FIGURE 11: Association rules for hydraulic supply line.

TABLE 9: Comparison of different minimum support thresholds.

	min_support	1-item set	2-item set	3-item set	4-item set	5-item set	6-item set	7-item set
Number of keywords	Multiple	1284	1071	665	308	152	40	15
	2	1284	1071	667	327	158	40	15
	5	634	834	142	1	—	—	—
	10	313	240	18	—	—	—	—
Number of items	Multiple	1284	3421	1632	445	130	35	8
	2	1284	4664	2770	653	159	42	9
	5	634	384	80	4	—	—	—
	10	313	165	20	—	—	—	—
Ratio	Multiple	100%	31%	41%	69%	117%	114%	188%
	2	100%	23%	24%	50%	99%	95%	167%
	5	100%	217%	178%	25%	—	—	—
	10	100%	145%	90%	—	—	—	—

FIGURE 12: Comparison of different minimum support thresholds.

FIGURE 13: Efficiency improvement.

5. Conclusions

In this article, a data-driven risk analysis approach is proposed to help superintendents identify and analyse hazards. Through the NLP and the improved Eclat algorithm, the historical hazard records are transformed into structured association rules. With the aid of change pattern mining, the hazard change trend can be identified. In actual work, as long as the superintendent delimitates the object, the relevant risk information such as potential hazards, conditional probabilities, and change trends can be automatically output. Based on this information, the pertinence of hazard identification and the efficiency of a hazard analysis can be improved. Case studies have shown that the efficiency of the improved algorithm is increased by 13.68%, and 59.66% of potential hazards can be identified in advance.

It should be pointed out that the multiple minimum support thresholds, minimum confidence threshold, and sliding window size are selected based on experiments. In actual work, superintendents can try to make adjustments

according to needs. Moreover, this approach requires a large amount of historical records to support data mining. The more historical records, the more abundant risk information will be output. Therefore, it is very essential to establish a systematic historical hazard database and record the hazard data continuously.

It is very meaningful to use data mining techniques to support risk analysis in complex workplaces. Keyword extraction, synonym merging, redundant keyword deletion, and unidentified hazard prediction need further research to improve the effectiveness of hazard text analysis and the accuracy of hazard prediction. The information system construction of the hazard analysis approach should also become the focus of future research.

Data Availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This research was funded by The National Key Research and Development Program of China, grant no. 2019YFC0810701.

References

- [1] B. J. Seligmann, E. Németh, and I. T. Cameron, "A blended hazard identification methodology to support process diagnosis," *Journal of Loss Prevention in the Process Industries*, vol. 25, no. 4, pp. 746–759, 2012.
- [2] P. Xin, F. Khan, and S. Ahmed, "Dynamic hazard identification and scenario mapping using Bayesian network," *Process Safety and Environmental Protection*, vol. 105, pp. 143–155, 2017.
- [3] A. Albert, M. R. Hallowell, and B. M. Kleiner, "Enhancing construction hazard recognition and communication with energy-based cognitive mnemonics and safety meeting maturity model: multiple baseline study," *Journal of Construction Engineering and Management*, vol. 140, no. 2, Article ID 04013042, 2013.
- [4] Y. M. Goh and D. Chua, "Case-based reasoning approach to construction safety hazard identification: adaptation and utilization," *Journal of Construction Engineering and Management*, vol. 136, no. 2, pp. 170–178, 2009.
- [5] X. H. Wang, X. F. Huang, and Y. Luo, "Improving workplace hazard identification performance using data mining," *Journal of Construction Engineering and Management*, vol. 144, 2018.
- [6] R. L. Daft, "A dual-core model of organizational innovation," *Academy of Management Journal*, vol. 21, no. 2, pp. 193–210, 1978.
- [7] P. Baybutt, "A critique of the hazard and operability (HAZOP) study," *Journal of Loss Prevention in the Process Industries*, vol. 33, pp. 52–58, 2015.
- [8] M. Giardina and M. Morale, "Safety study of an LNG regasification plant using an FMECA and HAZOP integrated methodology," *Journal of Loss Prevention in the Process Industries*, vol. 35, pp. 35–45, 2015.
- [9] F. Yan and K. Xu, "A set pair analysis based layer of protection analysis and its application in quantitative risk assessment," *Journal of Loss Prevention in the Process Industries*, vol. 55, pp. 313–319, 2018.
- [10] P. A. García Aguirre, L. Pérez-Domínguez, and D. Luviano-Cruz, "PFDA-FMEA, an integrated method improving FMEA assessment in product design," *Applied Sciences*, vol. 11, pp. 1–15, Article ID 1406, 2021.
- [11] A. C. Torres-Echeverria, "On the use of LOPA and risk graphs for SIL determination," *Journal of Loss Prevention in the Process Industries*, vol. 41, pp. 333–343, 2016.
- [12] B. Esmaili and M. R. Hallowell, "Diffusion of safety innovations in the construction industry," *Journal of Construction Engineering and Management*, vol. 138, no. 8, pp. 955–963, 2011.
- [13] M. Behm and A. Schneller, "Application of the Loughborough construction accident causation model: a framework for organizational learning," *Construction Management & Economics*, vol. 31, no. 6, pp. 580–595, 2013.
- [14] X. Ding, Q. Huang, and H. Zhu, "Subway tunnel disease associations mining based on fault tree analysis algorithm," *Mathematical Problems in Engineering*, vol. 2017, Article ID 2621493, 7 pages, 2017.
- [15] N. Hfaiedh, S. Kabiche, C. Delescluse et al., "Performing a preliminary hazard analysis applied to administration of injectable drugs to infants," *Journal of Evaluation in Clinical Practice*, vol. 23, no. 4, pp. 875–881, 2017.
- [16] H. Jeevith and R. Børge, "Applications of machine learning methods for engineering risk assessment—a review," *Safety Science*, vol. 122, pp. 1–30, 2019.
- [17] A. J.-P. Tixier, M. R. Hallowell, B. Rajagopalan, and D. Bowman, "Construction safety clash detection: identifying safety incompatibilities among fundamental attributes using data mining," *Automation in Construction*, vol. 74, pp. 39–54, 2017.
- [18] D. E. Brown, "Text mining the contributors to rail accidents," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 2, pp. 346–355, 2016.
- [19] L. Yang, K. Li, and G. Song, "Dynamic railway derailment risk analysis with text-data-based Bayesian network," *Applied Sciences*, vol. 11, Article ID 994, 2021.
- [20] H. R. Marucci-Wellman, H. L. Corns, and M. R. Lehto, "Classifying injury narratives of large administrative databases for surveillance-A practical approach combining machine learning ensembles and human review," *Accident Analysis & Prevention*, vol. 98, pp. 359–371, 2017.
- [21] N. G. Mutlu and S. Altuntas, "Assessment of occupational risks in Turkish manufacturing systems with data-driven models," *Journal of Manufacturing Systems*, vol. 53, pp. 169–182, 2019.
- [22] N. G. Mutlu and S. Altuntas, "Risk analysis for occupational safety and health in the textile industry: integration of FMEA, FTA, and BIFPET methods," *International Journal of Industrial Ergonomics*, vol. 72, pp. 222–240, 2019.
- [23] N. G. Mutlu and S. Altuntas, "Hazard and risk analysis for ring spinning yarn production process by integrated FTA-FMEA approach," *Textile and Apparel*, vol. 29, pp. 208–218, 2019.
- [24] S. D. Robinson, W. J. Irwin, T. K. Kelly, and X. O. Wu, "Application of machine learning to mapping primary causal factors in self reported safety narratives," *Safety Science*, vol. 75, pp. 118–129, 2015.
- [25] Y. J. Lu, F. Y. Tung, and C. H. Wang, "Developing an expert prognosis system of the reciprocating compressor based on associations among monitoring parameters and maintenance records," *Journal of Loss Prevention in the Process Industries*, vol. 69, no. 8, Article ID 104382, 2020.
- [26] C. Xu, J. Bao, C. Wang, and P. Liu, "Association rule analysis of factors contributing to extraordinarily severe traffic crashes in China," *Journal of Safety Research*, vol. 67, pp. 65–75, 2018.
- [27] Y. Song, S. Kou, and C. Wang, "Modelling crash severity by considering risk indicators of driver and roadway: a Bayesian network approach," *Journal of Safety Research*, vol. 76, pp. 64–72, 2020.
- [28] J. Wang, X. Pan, and L. Wang, "Method of spare parts prediction models evaluation based on grey comprehensive correlation degree and association rules mining: a case study in aviation," *Mathematical Problems in Engineering*, vol. 2018, Article ID 2643405, 10 pages, 2018.
- [29] S. Ma, X. Wang, and X. Wang, "A framework for diagnosing urban rail train turn-back faults based on rules and algorithms," *Applied Sciences*, vol. 11, Article ID 3347, 2021.
- [30] H. Salmane, L. Khoudour, and Y. Ruichek, "A video-analysis-based railway-road safety system for detecting hazard

- situations at level crossings,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, pp. 596–609, 2015.
- [31] Y. M. Goh, C. U. Ubeynarayana, K. L. X. Wong, and B. H. W. Guo, “Factors influencing unsafe behaviors: a supervised learning approach,” *Accident Analysis & Prevention*, vol. 118, pp. 77–85, 2018.
 - [32] R. Agrawal, H. Road, and S. Jose, “Mining association rules between sets of items in large databases,” in *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, May 1993.
 - [33] H. S. Song, J. K. Kim, and S. H. Kim, “Mining the change of customer behavior in an internet shopping mall,” *Expert Systems with Applications*, vol. 21, no. 3, pp. 157–168, 2001.
 - [34] M. Chen, A. Chiu, and H. Chang, “Mining changes in customer behavior in retail marketing,” *Expert Systems with Applications*, vol. 28, no. 4, pp. 773–781, 2005.
 - [35] M. Datar, A. Gionis, P. Indyk, and R. Motwani, “Maintaining stream statistics over sliding windows,” *SIAM Journal on Computing*, vol. 31, no. 6, pp. 1794–1813, 2002.
 - [36] M. J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li, “Parallel algorithms for discovery of association rules,” *Data Mining and Knowledge Discovery*, vol. 1, no. 4, pp. 343–373, 1997.
 - [37] J. Gehrke, F. Korn, and D. Srivastava, “On computing correlated aggregates over continual data streams,” *ACM SIGMOD Record*, vol. 30, no. 2, pp. 13–24, 2001.