BMC Public Health

RESEARCH Open Access

A chemical accident cause text mining method based on improved accident triangle

Zheng Li[1*], Min Yao[1,2], Zhenmin Luo[1], Xinping Wang[3], Tongshuang Liu[1], Qianrui Huang[1] and Chang Su[1]

Abstract

Background With the rapid development of China's chemical industry, although researchers have developed many methods in the field of chemical safety, the situation of chemical safety in China is still not optimistic. How to prevent accidents has always been the focus of scholars' attention.

Methods Based on the characteristics of chemical enterprises and the Heinrich accident triangle, this paper developed the organizational-level accident triangle, which divides accidents into group-level, unit-level, and workshop-level accidents. Based on 484 accident records of a large chemical enterprise in China, the Spearman correlation coefficient was used to analyze the rationality of accident classification and the occurrence rules of accidents at different levels. In addition, this paper used TF-IDF and K-means algorithms to extract keywords and perform text clustering analysis for accidents at different levels based on accident classification. The risk factors of each accident cluster were further analyzed, and improvement measures were proposed for the sample enterprises.

Results The results show that reducing unit-level accidents can prevent group-level accidents. The accidents of the sample enterprises are mainly personal injury accidents, production accidents, environmental pollution accidents, and quality accidents. The leading causes of personal injury accidents are employees' unsafe behaviors, such as poor safety awareness, non-standard operation, illegal operation, untimely communication, etc. The leading causes of production accidents, environmental pollution accidents, and quality accidents include the unsafe state of materials, such as equipment damage, pipeline leakage, short-circuiting, excessive fluctuation of process parameters, etc.

Conclusion Compared with the traditional accident classification method, the accident triangle proposed in this paper based on the organizational level dramatically reduces the differences between accidents, helps enterprises quickly identify risk factors, and prevents accidents. This method can effectively prevent accidents and provide helpful guidance for the safety management of chemical enterprises.

Keywords Accident triangle, Accident classification, Risk factors, Text mining, K-means algorithms

Background

China's five most dangerous industries are coal mining, metal and non-metal mining, construction, and chemical and fireworks manufacturing [1, 2]. Especially in the chemical industry, major accidents occur more frequently. According to statistics, from 2016 to 2021, there were 1050 chemical accidents in China, resulting in 1330 deaths. These major accidents lead to severe casualties, economic losses, environmental pollution, and other consequences and substantially affect society's

*Correspondence:
Zheng Li
zhengl0610@163.com
[1] College of Safety Science and Engineering, Xi'an University of Science and Technology, Xi'an 710054, China
[2] Institute of Management Science, Ningxia University, Yin'chuan 750021, China
[3] College of Management, Xi'an University of Science and Technology, Xi'an 710054, China

harmonious and stable development. Scholars have been studying how to avoid accidents for a long time [3, 4].

Heinrich[5] collected many industrial accident records from insurance companies through statistical analysis of many casualties. He proposed the accident triangle: the ratio of death, serious injury, minor injury, and non-injurious accidents in enterprises is 1:29:300. The accident triangle represents there are similar risk factors behind serious accidents and minor accidents, which have been used to guide the safety management of enterprises for a long time [6, 7]. Many sectors, such as the railway sector in the UK [8], the industrial sector in Germany [9], and the mining sector in Australia [10], conduct analyses of near misses and accidents without significant consequences to reveal operator errors and system deficiencies. It is generally believed that when the number of minor accidents increases, the accident triangle predicts that the number of serious accidents also increases, which can encourage enterprises to enhance safety management to prevent major accidents effectively[11].

Based on Heinrich's research findings, researchers from different engineering fields have also developed similar accident triangles [12]. The Bird accident triangle indicates that the ratio of serious or disabling injury, minor injuries, property damage accidents, and incidents with no visible injury or damage is 1:10:30:600 [13]. Tye-Pearson's principle states that the ratio of fatal or serious injury, minor injuries, first-aid treatment injuries, property damage accidents, and narrowly avoided accidents is 1:3:50:80:400 [14]. The International Association of Oil and Gas Producers (OGP) has collected the safety accident data since 1985. About 50 members of the oil and gas organization participated in the annual benchmark testing process, continuously updating their accident triangles [15].

However, with the deepening of research, some scholars questioned the effectiveness of the Heinrich accident triangle [16]. It mainly includes the following two points: (1) the ratio of lower to higher severity accidents exists in the form of a "safety-triangle"; (2) similar causes underlie both high and low severity events [17].

Regarding the first criticism, Rebbit suggested that, given that Heinrich's original data are not available, it is not possible to "verify or categorically refute" the specific ratio within the triangle. He further argued that the general way in which Heinrich categorized the safety accidents demarcated by severity (i.e., major, minor, and incident) makes it challenging to conduct replicable studies [18]. Marshall et al. stated that an analysis of occupational accidents across all the industries in Chile over 28-months shows that the ratio of fatal, serious and minor workplace accidents do not follow the ratio of accidents described in Heinrich accident triangle [19]. Yorio et al. studied the accident data from mines in the United States to confirm the predictive validity of the Heinrich accident triangle by checking if a certain number of accidents at a mine will produce a corresponding number of fatalities at the same mine. The results of their study did not match the figures described by Heinrich in his pyramid of accidents [17].

Regarding the second criticism, several studies have raised doubts about the assumption of similar causes, which determines whether enterprises focusing on minor accidents can effectively reduce the occurrence of major accidents [20–22]. Manuele argued that fatality and severe injury events often occur without any prior evidence or forewarning obtained through the analysis of less severe and near miss accidents [16]. Hale argued that the accident triangle has been abused and that preventing minor accidents will not automatically reduce serious accidents [23]. More and more evidence shows that companies with a very low incidence of minor accidents will also encounter serious accidents [24]. In the United States, although the incidence of non-fatal accidents in the entire workforce has decreased by 51% over the past decade, fatal accidents have only decreased by 25.5% [25]. Although studies have supported the notion that safety accidents delineated by degree have distinct causes [26] others found consistent causes between low and high severity events [8].

The traditional accident triangle usually analyzes personal injury accidents [27], dividing the accidents into death, serious injury, minor injury, and non-injury accidents for risk factor analysis. It rarely analyses property loss and environmental protection accidents, failing to identify some hidden risk factors [28]. The accident classification is based mainly on the consequences and types of accidents. The accident gradation is based on the accident severity. Its purpose is to provide a basis for preparing accident investigation reports, handling the person responsible for the accident, legal compensation, and other matters [29]. According to the "Byelaw Governing Reporting, Investigation and Handling of Production Safety Accidents" in China, accidents are divided into particularly serious accident, major accident, serious accident, and accident of minor seriousness in terms of bodily injuries and deaths or direct economic losses resulted in by the production safety accidents [30]. Based on the accident triangle, Sinopec proposes a three-level accident classification model from the perspective of organizational structure: group company-level accidents, recorded accidents, and accidents that require upgraded management. The accident gradation is more in-depth and detailed, and the requirements for casualties and direct economic losses are more strict [29].

The accident triangle is an adequate rule-of-thumb for safety planning; like any theory, it must be tested and updated. Industries need a set taxonomy for hazards/accidents and data to calibrate it [31]. Chemical enterprises are characterized by large scale, multiple organizational departments and levels, harsh production process conditions, complex equipment, and numerous risk factors [32, 33]. It is necessary to conduct a comprehensive risk factor analysis for chemical enterprises. Given the highly influential and debate of the accident triangle, additional research on the topic is important. This paper will test and update the accident triangle by developing a new accident classification method.

Although the industries differ, the accidents have similar trajectories [34]. Learning from the accident is considered the critical link to preventing future injuries [35], focusing on determining the root cause of the accident [36]. Currently, this work mainly depends on the judgment of domain experts, which is subjective and time-consuming. Enterprises have accumulated many safety accident reports. These unstructured text forms increase the difficulty of tacit mining knowledge. In recent years, data analysis in accident reports has provided a new way to research the causes of accidents [37]. As a branch of data analysis, text mining can extract unknown but valuable information and knowledge from unstructured text sets, involving knowledge in multiple fields such as artificial intelligence, machine learning, and natural language processing (NLP) [38]. It is currently a research hotspot in text information processing [39]. Since Feldman et al. [40] first proposed the concept of text mining in 1995, the development of text mining technology has become mature. It has been widely applied in fields such as biomedical [41], consumer behavior [42], emotional analysis [43], coal mining production [44, 45], transportation [46], and construction [47].

In the field of engineering safety management, the dispersion, diversity, and massive nature of safety data have led to difficulties in collecting and processing safety texts, thereby promoting the application of text mining technology [48]. In recent years, a few scholars have utilized text mining technology to extract key accident features and risk factors from accident investigation reports, fully leveraging the role of accident investigation reports in summarizing experiences and lessons learned and curbing accident risks [49, 50]. For example, Gao et al. [51] developed a verb-based text mining method that extracted the causes and results of 945 car traffic accident reports, which helps to understand the true causes of traffic accidents. Qiu et al. [52] combined text mining technology with complex networks to explore the causal mechanisms of coal mine accidents. Through text mining of 307 accident reports, 52 main accident causal factors were identified, and a coal mine accident causal network was constructed based on strong association rules between factors, providing a new perspective for identifying accident causes and their complex interaction mechanisms from accident report data. Esmaeili et al. [53] used text mining technology to obtain accident attribute feature values from over 1000 construction accident reports and conducted statistical analysis, ultimately identifying the risk factors of construction accidents. Raviv et al. [54] analyzed 212 near-miss and accident reports on tower cranes using text mining and k-means algorithms and found that technical failures are the most dangerous risk factor in tower cranes.

In the field of chemical safety management, work on anomaly detection [55], ontology-based knowledge acquisition [56], and process alarm prediction [57] have been undertaken based on accident texts. Despite such work, no existing method meets the demands of both universality and accuracy, and there still needs to be an efficient, convenient universal tool for extracting risk factors from chemical accident cases.

To solve the problem that serious accidents are difficult to eradicate, we select chemical enterprises with many risk factors for empirical research to explore the factors causing severe accidents. Firstly, based on the accident triangle, combined with the characteristics of large-scale, multiple organizational departments and levels of the chemical industry, the organizational-level accident triangle has been proposed. Try to determine the distribution law of risk factors in chemical enterprises through a new classification method. Secondly, this paper uses text mining technology to conduct text clustering analysis on unstructured accident reports, quickly find the risk factors behind different accidents, and solve the disadvantages of time-consuming and labor-consuming finding accident causes in the past.

Methods

This study comprehensively considers personal injury accidents, property loss accidents, and environmental accidents in enterprises. It proposes an accident triangle based on the organizational level, which classifies accidents into group-level accidents, unit-level accidents, and workshop-level accidents. The collected accident reports are model input, and each accident is classified into levels using the organizational level accident triangle. Then, text mining technology is used to cluster and analyze the causes of accidents at different organizational levels by identifying the various risk factors contained in accidents at different organizational levels, namely the model output, to help enterprises formulate targeted risk control measures. The overall flow of this model is shown in Fig. 1.

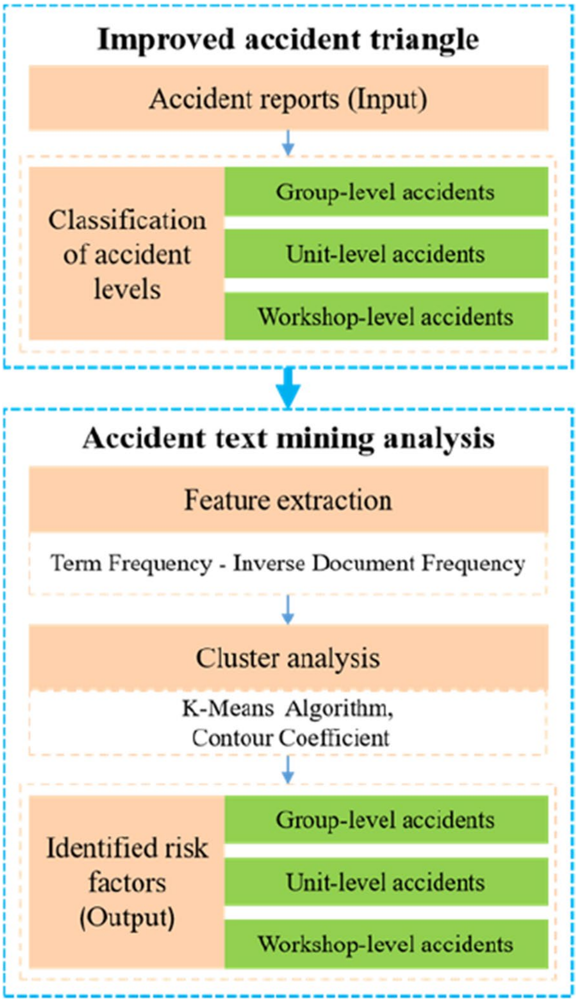

Improved accident triangle
Accident reports (Input)
Classification of accident levels
Group-level accidents
Unit-level accidents
Workshop-level accidents
Accident text mining analysis
Feature extraction
Term Frequency - Inverse Document Frequency
Cluster analysis
K-Means Algorithm, Contour Coefficient
Identified risk factors (Output)
Group-level accidents
Unit-level accidents
Workshop-level accidents

Fig. 1 Overall Flowchart of the proposed model

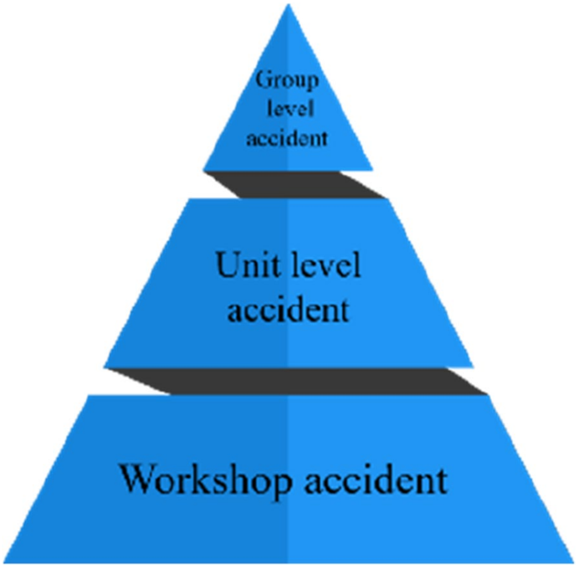


Fig. 2 The organization-level accident triangle

Improved accident triangle

The traditional accident triangle usually analyzes personal injury accidents and classifies the accident severity into four categories: death, serious injury, minor injury, and non-injury. However, enterprises will also have accidents of different natures in daily production, such as production, quality, environmental, and property loss accidents [58]. This study will propose a more comprehensive accident classification method that divides accidents into personal injury accidents, property loss accidents, and environmental protection accidents, addressing the drawbacks of traditional accident triangles that cannot analyze all risk factors. For example, traditional accident triangles focus on analyzing personal injury accidents, often categorizing property loss accidents and environmental accidents roughly as non-injury accidents or near misses, which is not conducive to further accurately identifying the risk factors that cause accidents.

This study focuses on identifying accident risk factors in China's chemical industry, improving the accident triangle, and reclassifying accidents from the organizational level. The organization-level accident triangle divides enterprise accidents into group-level, unit-level, and workshop-level. In risk factor mining for different accident levels, analyze the ratio and causes of different accident levels. Accidents at all levels are classified according to severity; the definitions and classification standards for group-level, unit-level, and workshop-level accidents are shown in Additional file 1. The bottom of the organization-level accident triangle is the number of workshop-level accidents; the middle refers to unit-level accidents; the top is the number of group-level accidents (Fig. 2).

Applying the organizational-level accident triangle includes the following processes: In Step 1, the relevant information of three levels of accidents is counted, and the accident management account is updated in time. Step 2, whether the accident triangle based on the organization level is in the state of a "positive triangle" or "inverted triangle" is checked by counting the frequency of occurrence of three levels of accidents. Accidents at different levels are divided according to severity. Based on practical experience, there are far more minor accidents than serious ones. Then, the number of workshop-level accidents should be the largest, the number of unit-level accidents should rank second, and the number of group-level accidents should be the least. If the proportion of accidents at different levels meets this requirement, it is considered in the positive triangle state. Otherwise, it is in the inverted triangle state. Step 3, if in the "positive triangle" state, indicates that the design of enterprise safety rules and regulations is reasonable. Further, it analyzes various accidents, determines safety management loopholes, and eliminates relevant risk factors. Otherwise, it indicates that the design and implementation of enterprise safety management rules and regulations are unreasonable. The enterprise needs to solve this problem and use text mining technology to ascertain the causes of accidents, reveal management loopholes, and eliminate risk factors. Step 4 is accident statistics for the enterprise again, followed by the second step to start a new round of analysis. The process is shown in Fig. 3.

This paper uses the Pearson correlation coefficient method to analyze the correlation among organizational-level, unit-level, and workshop-level accidents. The correlation coefficient is expressed in R, with a value range of 0 to 1.

$$R = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left[n\sum X^2 - \left(\sum X\right)^2\right]\left[n\sum Y^2 - \left(\sum Y\right)^2\right]}} \tag{1}$$

Where, n refers to the total number of accidents, while X and Y refer to the number and average value of accidents at different levels. The closer the absolute value of R is to 1, the better the correlation.

Text mining analysis of accidents at different organizational levels

Text mining technology has been widely used to study accident causes. Through extensive research of accident reports, text mining can better understand the causes of accidents and significantly improve the accuracy of accident prediction [59].

In recent years, the analysis of incident/accident data using data mining techniques and algorithms has attracted much attention among researchers [60, 61], which promotes the birth and development of data-mining technology [62]. Many data mining technologies, such as Support Vector Machine [63], classification analysis [64], cluster analysis [65], association analysis [66], chi-square automatic interaction detection [67], and Bayesian networks [68], are used to identify the hidden patterns and structures in the safety database. This article uses the text clustering method to identify risk factors in accident text, and the analysis process is shown in Fig. 4.

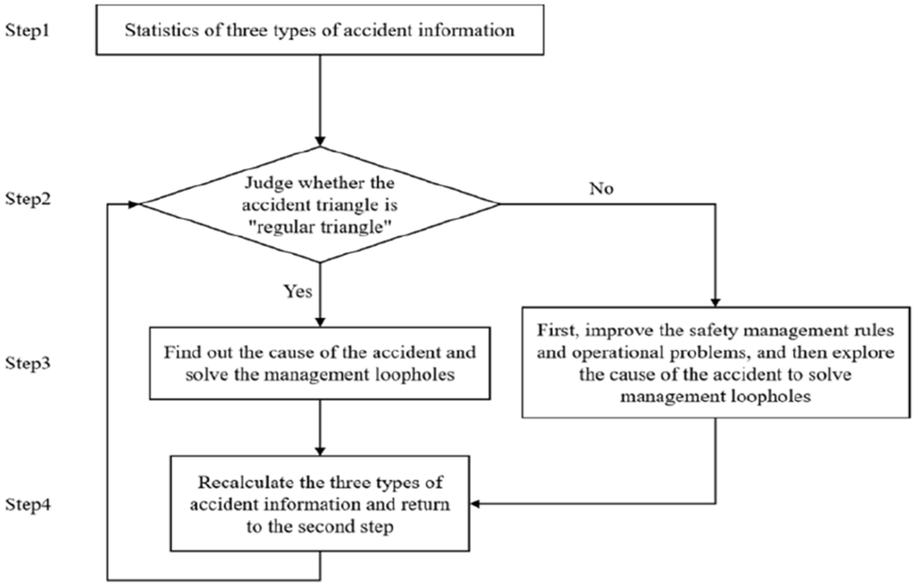

Step1
Statistics of three types of accident information
Step2
Judge whether the accident triangle is "regular triangle"
No
Yes
Step3
Find out the cause of the accident and solve the management loopholes
First, improve the safety management rules and operational problems, and then explore the cause of the accident to solve management loopholes
Step4
Recalculate the three types of accident information and return to the second step

Fig. 3 Analysis flowchart for the use of the organizational-level accident triangle

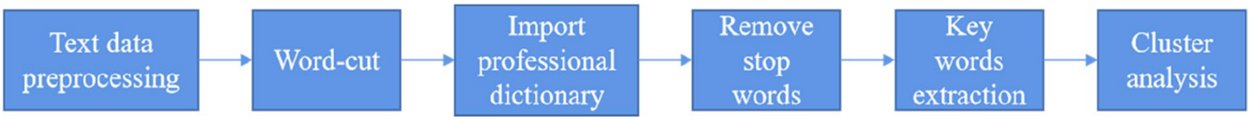


Fig. 4 Text mining analysis process

Extract keywords

Text keyword extraction requires pre-processing technology to convert the text into a form the computer can recognize. This paper uses the Jieba Chinese word segmentation tool to segment and label accident records. To identify professional terms and idioms and ensure that these words are not segmented, compiled a professional dictionary according to the vocabulary of the coal chemical industry and the expression characteristics of enterprise safety officers on potential safety hazards and accident records. Stop words are words and tonal symbols that frequently appear in the text but have no functional meaning and do not help analyze the main idea of the text. These meaningless words can be deleted by importing the stop word list. After getting keywords and assigning weight to each keyword, the Term Frequency Inverse Document Frequency (TF-IDF) is usually used as a feature evaluation function for feature extraction [69]. TF-IDF is expressed as:

$$TF - IDF_{i,j} = tf_{i,j} \times idf_{i} \tag{2}$$

$$tf_{i,j} = \frac{n_{i,j}}{\sum_{K} n_{k,j}} \tag{3}$$

$$idf_{i} = \log\frac{|D|}{1 + |\{j : t_i \in d_j\}|} \tag{4}$$

Where, $n_{i,j}$ denotes the number of occurrences of the keyword t_i that appears in the accident record document d_j, and $\sum_K n_{k,j}$ is the number of all keywords in the accident record document d_j $|D|$ represents the total number of accident record documents, and $\left|\{j : t_i \in d_j\}\right|$ is the number of documents containing keyword t_i, to avoid this item being zero and the divisor being zero, it is generally expressed as $1 + \left|\{j : t_i \in d_j\}\right|$.

TF denotes the number or frequency of a word in the article. The word is essential if a keyword appears multiple times in an accident record document. IDF represents the recognition degree of a keyword in the accident record document. The larger IDF value means the keyword is essential in this document and vice versa. TF-IDF integrates the advantages of TF and IDF [70].

Text clustering analysis

Clustering is widely applied in machine learning and data mining as a standard data research method. The common clustering methods are the minimum distance within the group and the maximum distance between groups [71]. The basic idea of the *K*-means algorithm proposed by MacQueen [72] is to divide all objects into *K* clustering centers according to the nearest principle. The similarity between texts is measured by Euclidean distance. Before *K*-means clustering, we need to determine the K value and the number of clusters. Randomly obtain *K* initial cluster centers and iterate the average similarity between documents until the optimal solution is derived.

K-means clustering algorithm takes the sum of squares errors (SSEs) as the objective function to minimize the SSEs between texts in *K* clusters. The cluster center e_i of cluster E_i can be expressed as:

$$e_i = \frac{1}{n_i}\sum_{n \in E_I} x \tag{5}$$

The SSE between texts is calculated as follows:

$$SSE = \sum_{i=1}^{K}\sum_{n \in E_i} cos(e_i, x)^2 \tag{6}$$

Where, x represents the text object, E_i is the i^{th} cluster, n_i denotes the number of samples therein, and e_i is the center of cluster E_i.

Peter [73] first proposed the contour coefficient to calculate the *K* value in *K*-means text clustering, which can judge the text clustering effect. Averaging the contour coefficients of all vectors is the contour coefficient of the cluster. The contour coefficient's absolute value is not greater than 1. The larger the average value of the contour coefficient, the better the text clustering effect.

Results

Database

The data in this paper comes from a large coal chemical enterprise in China. The group has more than a dozen secondary units, such as a coking plant, methanol plant, olefin plant, and power company. The company has formulated safety management regulations by relevant national laws, such as "Classification Standard for Casualty Accidents of Enterprise Employees" (GB6441), "Identification of Labor Ability and Disability Grade of Employees Caused by Industrial Injury and Occupational Disease" (GB/T16180), "Byelaw Governing Reporting, Investigation and Handling of Production Safety Accidents" (An order by PRC State Council No. 493), and conducted daily safety management and accident investigation under regulations.

The workflow for accident investigation is shown in Fig. 5. During the accident investigation, the safety technician and safety supervisor are the backbone of promoting the smooth progress of the accident investigation. To ensure the objectivity and authenticity of the accident report, the responsible person related to the accident cannot be a member of the accident investigation team. After the accident investigation, the safety department

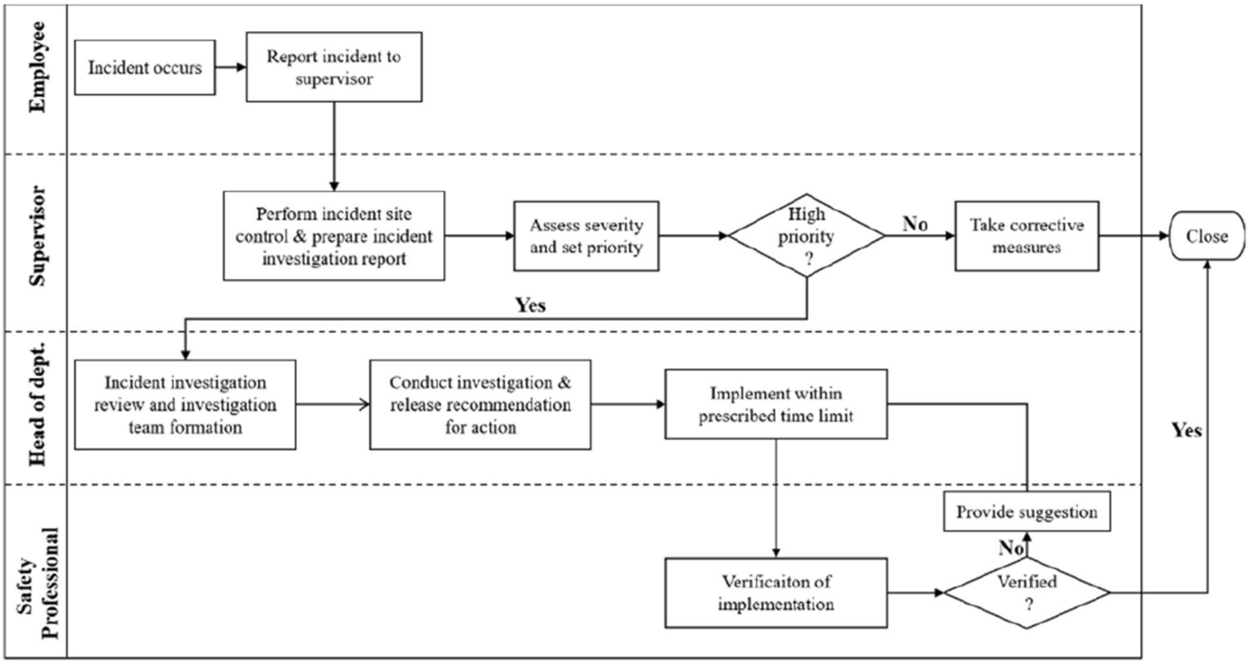

Employee
Incident occurs
Report incident to supervisor
Supervisor
Perform incident site control & prepare incident investigation report
Assess severity and set priority
High priority ?
No
Take corrective measures
Close
Yes
Head of dept.
Incident investigation review and investigation team formation
Conduct investigation & release recommendation for action
Implement within prescribed time limit
Yes
Safety Professional
Provide suggestion
No
Verificaiton of implementation
Verified ?

Fig. 5 Proposed workflow through the accident investigation module

will prepare the accident report as required according to the collected data.

The database used in this paper consists of 484 accident records generated by SMS from 2015 to 2020; Additional file 2 lists information from two of the accident reports. The accident records contain all accident information of the enterprise, including accident position, accident time, accident level, accident nature, consequence degree, and accident process.

The database used in the work presented here records the accident information of the sample enterprises in a certain period of actual production. Because all secondary units conducted daily safety management under the safety management rules and regulations formulated by the group safety department, all safety managers and on-site operators unified the rules and regulations. The accident data can reflect various risk factors in enterprises, and it is feasible to use the method proposed in this paper to identify enterprise risk factors.

Correlation analysis of accidents at different organizational levels

A statistical analysis of the company's accidents from 2015 to 2020 is conducted from the organizational level perspective (Table 1). The ratio of group-level, unit-level, and workshop-level accidents is 1:3:12; this division dramatically reduces the difference between accidents and helps find common causes at different organizational levels.

Table 1 Annual accident statistics of sample enterprises at different organizational levels

Year	Group-level	Unit-level	Workshop-level
2015	5	23	60
2016	4	11	79
2017	4	11	70
2018	5	11	67
2019	6	12	51
2020	7	12	46

To more intuitively display the annual change trend of accidents at different organizational levels, a trend chart of accident rate change has been prepared, as shown in Fig. 6. The trend of change in unit-level and group-level accident rates is highly similar. In contrast, the changing trend in both is opposite to that in the workshop-level accident rate. Spearman correlation analysis is performed among the group, unit, and workshop accident rates, and the results are shown in Table 2. The correlation coefficient between the group-level accident rate and the unit-level accident rate is 0.657, which passes the significance test at 0.05,, showing a significant positive correlation between the unit-level accident rate and group-level accident rate. In enterprise safety management, rectification measures should be implemented for unit-level accidents. Reducing the unit accident rate can also reduce group-level accidents and overcome the difficulty of

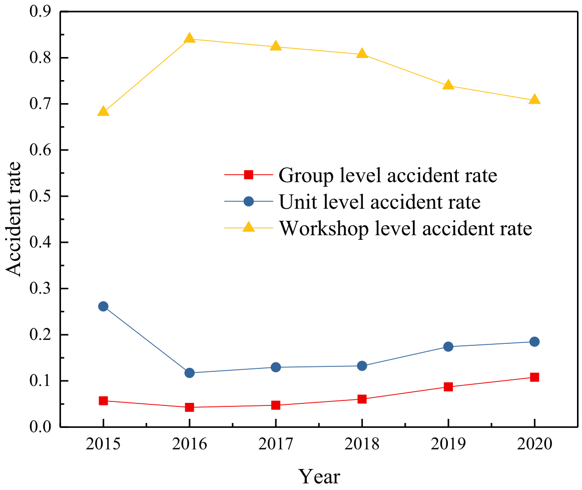


Fig. 6 Trends of accident rate at different organizational levels

finding causal factors because the number of group-level accidents is too small. The group-level and unit-level accident rates negatively correlate with the workshop-level accident rates; the correlation coefficients are -0.594 and -0.857, respectively, and pass the significance test at the level of 0.05. The enterprise accident data are divided into different levels of accidents; when the proportion of group-level and unit-level accidents is relatively small, the proportion of workshop-level accidents is relatively large. On the other hand, it can also be explained that when workshop-level accidents occur frequently, staff and managers will spend more time paying attention to safety problems and constantly solve the neglected risks arising from minor accidents to avoid more severe accidents.

Text mining analysis

The work divides enterprise accidents into group, unit, and workshop levels based on the organizational-level accident triangle. The characteristics and risk factors of the three levels of accidents will be found by text mining. This paper will conduct text mining on 484 accident records, including 373 accident records at the workshop level, 80 at the unit level, and 31 at the group level.

Extract keywords

The TF-IDF method assigns weight to accident keywords at all levels. The TF-IDF values of each keyword are arranged in descending order, as illustrated in Table 3, 4 and 5. The top 30 accident keywords at all levels are displayed. The keywords with a high weight of workshop-level accidents include gasifier, boiler, inspection,

Table 2 Correlation analysis of accident rates at different organizational levels

		Group-level accident rate	Unit-level accident rate	Workshop accident rate
Group-level accident rate	Correlation coefficient	1		
	Significance	0		
Unit-level accident rate	Correlation coefficient	0.657	1	
	Significance	0.038	0	
Workshop accident rate	Correlation coefficient	-0.594	-0.857	1
	Significance	0.016	0.008	0

Table 3 Weighted values of keywords in workshop-level accident keywords (top 30)

No	Keywords	TF-IDF	No	Keywords	TF-IDF	No	Keywords	TF-IDF
1	Gasifier	0.1041	11	Flow	0.0472	21	Furnace shutdown	0.0308
2	Boiler	0.0715	12	Overhaul	0.0458	22	Surge	0.0307
3	Central Control Room	0.0698	13	Discover	0.0423	23	Gasification	0.0302
4	Load	0.0601	14	Interlock	0.0404	24	Scheduling	0.0293
5	Check	0.0590	15	Pressure	0.0386	25	Pulverized coal	0.0287
6	Induced draft fan	0.0589	16	Valve	0.0331	26	Meter	0.0283
7	Fan	0.0550	17	Start-up	0.0322	27	Compressor	0.0280
8	Running	0.0531	18	Trip	0.0322	28	Conversion	0.0280
9	Workshop	0.0518	19	Switch	0.0319	29	Shut down	0.0277
10	Liquid level	0.0518	20	Stop	0.0319	30	Economizer	0.0257

Table 4 Weighted values of keywords in unit-level accidents (top 30)

No	Keywords	TF-IDF	No	Keywords	TF-IDF	No	Keywords	TF-IDF
1	Scene	0.1000	11	Pipeline	0.0437	21	Display	0.0306
2	Central control room	0.0954	12	Compressor	0.0427	22	Stop	0.0304
3	Coke oven gas	0.0795	13	Load	0.0393	23	Pressure relief	0.0304
4	Interlock	0.0729	14	Liquid level	0.0390	24	Supercharger	0.0304
5	Gasifier	0.0604	15	Slag lock bucket	0.0376	25	Synthesis	0.0300
6	Scheduling	0.0544	16	Running	0.0372	26	Close	0.0300
7	Workshop	0.0497	17	Work	0.0371	27	Methanol	0.0296
8	Notice	0.0490	18	Pressure	0.0365	28	Main operation	0.0289
9	Induced draft fan	0.0447	19	Overhaul	0.0320	29	Water seal	0.0286
10	Reactor	0.0441	20	Flow	0.0313	30	Check	0.0285

Table 5 Weighted values of keywords in group-level accidents (top 30)

No	Keywords	TF-IDF	No	Keywords	TF-IDF	No	Keywords	TF-IDF
1	Reactor	0.1028	11	Workshop	0.0470	21	Fan	0.0357
2	Scene	0.0777	12	Be on duty	0.0454	22	Conversion	0.0356
3	Coke quenching car	0.0648	13	Central control room	0.0449	23	Boiler	0.0356
4	Pressure	0.0640	14	Rotating speed	0.0426	24	Electric machinery	0.0355
5	Liquid level	0.0623	15	High-pressure cylinder	0.0424	25	Coke quenching	0.0349
6	Main operation	0.0598	16	Stop	0.0399	26	Lean liquid pump	0.0349
7	Scheduling	0.0548	17	Pure benzene	0.0399	27	Surge	0.0330
8	Coking plant	0.0542	18	Feedwater pump	0.0377	28	Propylene	0.0321
9	Compressor	0.0532	19	Check	0.0365	29	Lubricating oil	0.0318
10	Vibration	0.0504	20	Loading	0.0362	30	Pipeline	0.0317

induced draft fan, compressor, economizer, liquid level, and tripping. It indicates that workshop-level accidents are mostly related to this equipment. The accident records demonstrate that the keywords with high weight ranking can accurately reflect the frequent workshop-level accidents.

The keywords with a high weight of unit-level accidents include coke oven gas, interlock, gasifier, induced draft fan, reactor, and pipeline, suggesting that unit-level accidents are mostly related to the above keywords. The accident records show that the causes can be classified as coke oven gas leakage, pipeline rupture, induced draft fan component damage, and excessive fluctuation of process parameters.

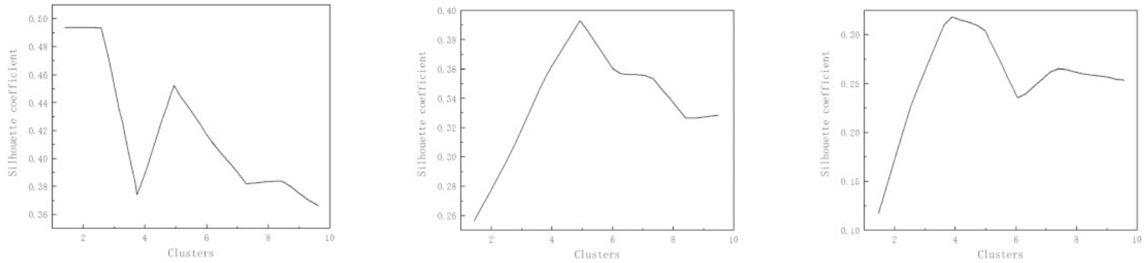
The keywords with high weights among group-level accident reports include reactor, coke quenching vehicle, pressure, liquid level, coking plant, compressor, vibration, and pure benzene, implying that group-level accidents are mostly related to the above keywords. By comparing the accident records, group-level accidents are found to be mainly death and severe injury accidents caused by toxic substance leakage, mechanical injury and safety barrier damage, environmental accidents caused by ammonia and raw gas leakage, shutdown accidents arising from equipment damage, and excessive fluctuation of process parameters.

Text clustering analysis

The contour coefficient is calculated for the text data of accident records at different levels (Fig. 7). The corresponding contour coefficient value is the largest when the K values of workshop-level, unit-level, and group-level accident clusters are 5, 5, and 4. The optimal number of clusters for K-means text clustering of workshop-level, unit-level, and group-level accidents are 5, 5, and 4, respectively.

Programming Python conducts K-means cluster analysis on accident records at different levels. The clustering result of workshop-level accidents is shown in Fig. 8; see Additional file 3 for detailed results.

The accident records in the five clusters are 181, 84, 41, 40, and 27, respectively. The accident type, high-frequency keywords, and accident cause analyses of each cluster are displayed in Table 6. The accident types of



(a) Workshop-level accident cluster contour coefficient diagram (b) Unit-level accident cluster contour coefficient diagram (c) Group-level accident clustering contour coefficient diagram

Fig. 7 Clustering contour coefficients of accidents at different levels

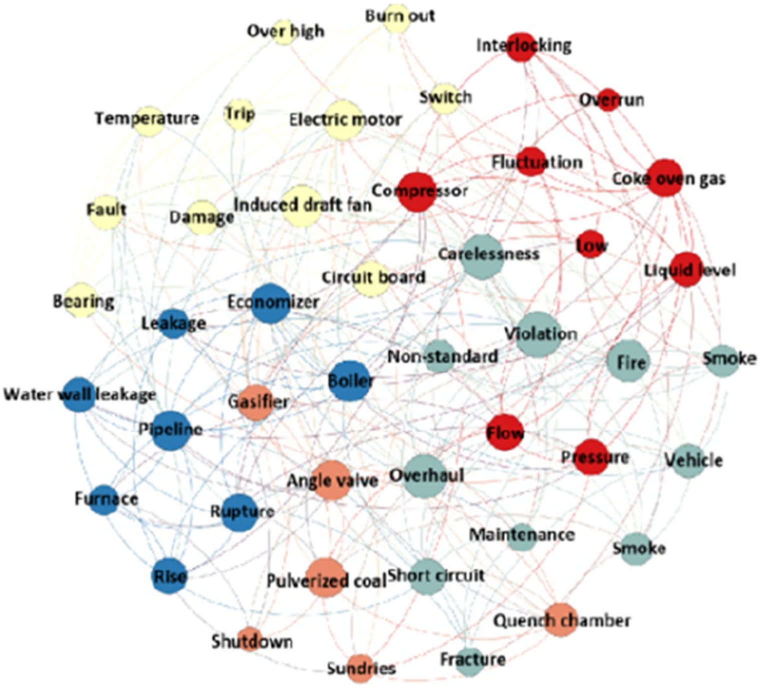


Fig. 8 Visual clustering result of workshop-level accidents

cluster 0 are mainly personal injury accidents; the main reason is unsafe personal behavior, and the unsafe state of things causes only a few accidents. Cluster 1 is mainly the production accident of shutdown and maintenance caused by the damage of induced draft fan, motor, and other equipment components. Cluster 2 primarily refers to shutdown and maintenance accidents caused by economizer leakage and pipeline rupture. Cluster 3 mainly represents shutdown accidents caused by excessive fluctuation of process parameter values and environmental accidents caused by hazardous substance leakage. Cluster 4 mainly refers to equipment-tripping accidents caused by untimely cleaning of sundries or malfunction, which also reflects the management vulnerability of the enterprise in terms of cleanliness and hygiene.

In the workshop-level accidents cluster, the proportions of personal injury, property loss, and environmental accidents are 29%, 69%, and 2%. The leading causes of personal injury accidents are weak safety awareness of employees, failure to take protective measures, illegal operation, or nonstandard operation. The leading causes of property loss accidents are fire caused by a short circuit, untimely cleaning of sundries, and interlocking shutdown caused by equipment failure, leakage, and excessive fluctuation of process parameters. The leading cause of environmental accidents is the excessive discharge of pollutants caused by equipment failure and leakage.

Table 6 Cluster analysis of workshop-level accidents

Cluster no	Accident types	High-frequency keyword	Main causes of accidents
0	Traffic accident, fire accident, mechanical injury accident, fall down accident, scald accident	Carelessness, vehicle, collision, fracture, fire, smoke, short circuit, crush injury, overhaul, violation, non-standard, maintenance, disassembly, cleaning, safety rope, slip, fall, high temperature, scald	Weak safety awareness, failure to take protective measures, non-standard operation, untimely communication, untimely equipment maintenance, circuit short circuit, and pipeline leakage
1	Production accident	Induced draft fan, electric motor, circuit board, bearing, switch, damage, fault, temperature, over high, burn out, trip, stop	Equipment failure, induced damage to draft fan parts, electric motor damage, over-temperature, switch trip
2	Production accident	Low temperature, high temperature, Economizer, boiler, pipeline, water wall leakage, rupture, leakage point, water level, rise, furnace, pressure, rise, boiler shutdown, maintenance	Economizer leak, pipe rupture, water wall rupture
3	Production accident, environmental pollution accident	Compressor, coke oven gas, flow, pressure, liquid level, too low, rise, fluctuation, interlocking, overrun, trip, shutdown, flare, black smoke, alarm	Significant fluctuation of gas flow/ liquid level/ pressure/ temperature, excessive vibration amplitude, damage to compressor components, pipe blockage, and flange leakage
4	Production accident	Angle valve, pulverized coal, gasifier, quench chamber, sundries, trip, shutdown	Untimely removal of sundries, malfunction, and excessive fluctuation of process parameters

The cluster result analysis of unit-level accidents is shown in Fig. 9, and accident cause analyses of each cluster are displayed in Table 7.

Cluster 0 mainly includes shutdown maintenance and environmental accidents caused by pipeline blockage and device damage. Cluster 1 mainly denotes personal injury accidents caused by employees' non-standard operation or failure to take protective measures. Cluster 2 mainly refers to shutdown accidents caused by power system faults such as short-circuiting and substation tripping. Cluster 3 mainly denotes traffic accidents. Cluster 4 mainly refers to equipment tripping accidents caused by excessive fluctuation of process parameters.

Among unit-level accidents, personal injury, property loss, and environmental accidents accounted for 23%, 61%, and 16%, respectively. The leading cause of personal injury accidents is the non-standard operation of employees. The leading causes of property loss accidents are equipment failure, excessive fluctuation of process parameters, pipeline blockage, and leakage. The leading causes of environmental accidents are the excessive

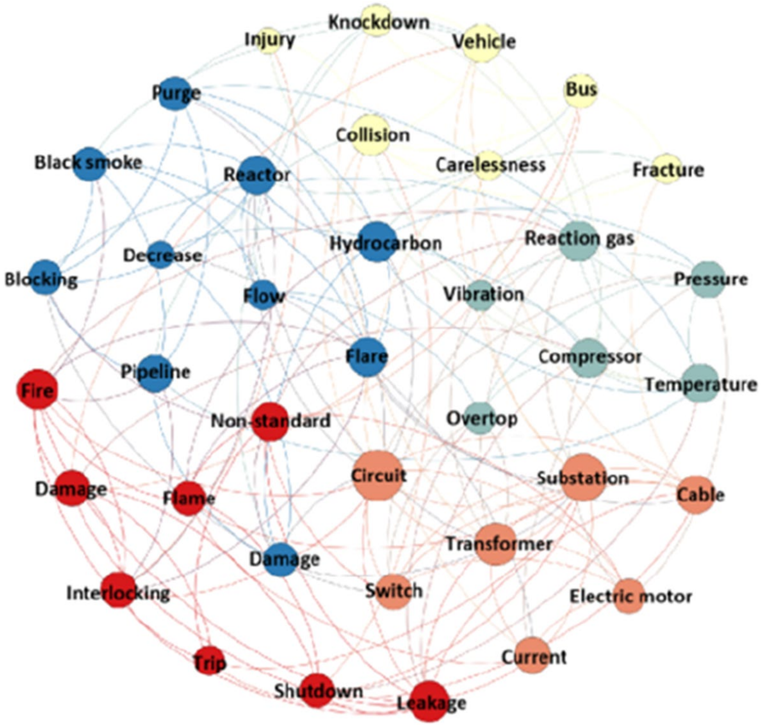


Fig. 9 Visual clustering result of unit-level accidents

Table 7 Cluster analysis of unit-level accidents

Cluster No	The accident types	High-frequency keyword	Main causes of accidents
0	Production accident, environmental pollution accident	Hydrocarbon, flare, black smoke, reactor, pipeline, purge, blocking, flow, liquid level, decrease, locking	Pipeline blockage, device damage
1	Fire accident, production accident, mechanical injury accident, scald accident	Non-standard, leakage, fire, smoke, flame, interlocking, shutdown, damage, trip	The operations are not standardized, and protective measures are not taken
2	Production accident	Circuit, substation, transformer, cable, switch, trip, current, high, electric motor	Short circuit, equipment power loss, voltage fluctuation, power trip
3	Traffic accident	Collision, vehicle, carelessness, bus, knockdown, fracture, injury	Lack of safety awareness and failure to comply with traffic rules
4	Production accident	Reaction gas, compressor, temperature, pressure, vibration, differential pressure, liquid level, overtop, too fast, interlock, trip	Excessive temperature / pressure / liquid level / vibration fluctuation

emission of harmful gases caused by substation failure, line short circuits, and control system failure.

The cluster result analysis of group-level accidents is shown in Fig. 10, and accident cause analyses of each cluster are displayed in Table 8.

Cluster 0 mainly refers to personal injury accidents caused by unsafe factors such as employees' weak safety awareness and illegal operations. Cluster 1 mainly denotes equipment shutdown accidents caused by excessive fluctuation of process parameters. Cluster 2 mainly indicates equipment tripping accidents caused by damage and short-circuiting. Cluster 3 mainly involves production accidents, but many causes include damage to equipment components, excessive fluctuation of process parameters, and untimely removal of sundries. The clustering effect is poor due to the small number of accidents.

In the group-level accidents cluster, personal injury, property loss, and environmental accidents accounted for 22%, 64%, and 14%, respectively. Personal injury accidents cause death and severe injury, and the main risk factors are illegal operations and failure to take protective measures. The main risk factor of property loss accidents is the failure of the compressor, separator, fan, and transformer. The main risk factor of environmental accidents is the leakage of pipes, flanges, and valves.

The clustering results show that the most significant proportion of accidents at the three levels is human factors, such as insufficient safety awareness of employees, illegal operations, and failure to inspect and repair equipment as required. It indicates that enterprises urgently need to train employees in safety knowledge and skills in future safety management. The leading causes of workshop-level accidents are equipment damage and pipeline leakage. It is necessary to conduct targeted maintenance and replacement of equipment components, such as induced draft fans, compressors, and economizers, and establish an assessment mechanism for regular maintenance. Power system faults account for a high proportion of unit-level and group-level accidents. Such accidents occur infrequently, but the degree of harm is enormous; it is necessary to strengthen the operation and maintenance of such power systems.

Practical application

Improvement measures

Based on the risk factors identified by cluster analysis results, the enterprise has rectified activities from the perspective of unsafe human behavior and the unsafe state of objects. Due to inadequate safety awareness, irregular operations, and failure to take proper protective measures among employees, the vast majority of accidents have occurred. The company has established a safety inspection team directly responsible for senior leaders, providing two months of safety knowledge and

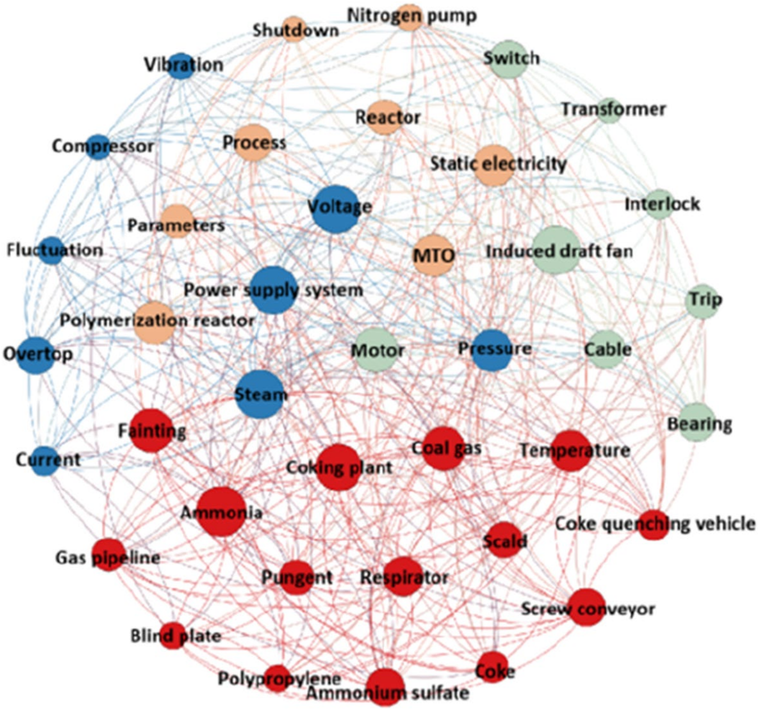


Fig. 10 Visual clustering result of group-level accidents

Table 8 Cluster analysis of group-level accidents

Cluster No	The accident types	High-frequency keyword	Main causes of accidents
0	Poisoning accident, explosion accident, mechanical injury accident, scald accident, environmental pollution accident, product quality accident	Coking plant, coal gas, ammonia, temperature, respirator, fainting, ammonium sulfate, coke quenching vehicle, scald, pungent, coke, screw conveyor, gas pipeline, blind plate, polypropylene, death	Weak safety awareness, failure to take protective measures, illegal operation, imperfect product quality management system
1	Production accident	Power supply system, voltage, steam, pressure, temperature, overtop, vibration, current, fluctuation, compressor, blockage	Excessive temperature/ pressure/current/vibration fluctuation and pipe network blockage
2	Production accident	Induced draft fan, motor, cable, bearing, temperature, trip, interlock, transformer	Equipment components damaged, short circuit, switch trip
3	Production accident	Polymerization reactor, static electricity, process parameters, MTO reactor, pressure, temperature, vibration, nitrogen pump, shutdown	Equipment components damaged, static electricity, excessive pressure/ temperature/vibration fluctuation, and accumulated materials are not cleaned timeously

technical training for all employees. The safety inspection team evaluates the training results of employees, and employees who fail the assessment are not allowed to take up their positions. In addition, an online knowledge learning and testing system has been designed based on different departments and job responsibilities. Employees can log in to the system through their mobile phones for problem-seeking and learning and must complete 8 h of learning tasks monthly. The department head randomly selects test questions from the system every month for testing, and the test results are included in the employee's performance evaluation.

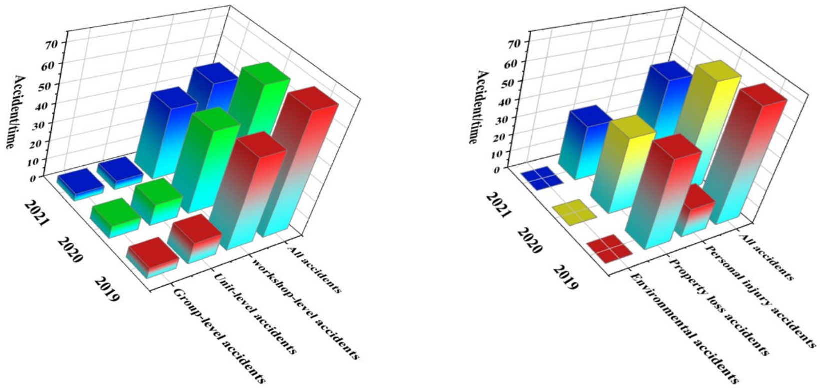
The enterprise has updated and maintained the frequently malfunctioning equipment and systems identified through cluster analysis of group-level, unit-level, and workshop-level accidents, such as induced draft fans, compressors, and power systems. In particular, the frequency of inspection and maintenance has been increased for critical equipment and components, and a comprehensive daily, weekly, monthly, and quarterly inspection and maintenance system has been developed. Finally, after addressing all local risks, the enterprise has developed standardized operation manuals and institutional norms for personnel, equipment, materials, and environmental factors involved in the production process, forming a standard model based on scientific and rigorous work processes and long-term accumulated experience and habits, forming a deeply rooted code of conduct in the minds of members, minimizing uncontrollable factors, and effectively helping the enterprise maintain a high level of safety.

Implementation effect

The safety situation of the enterprise has been dramatically improved after rectification measures. We have statistics on the accidents in 2021, as shown in Fig. 11. The accidents in 2021 have decreased by 29% and 25% compared with 2019 and 2020. Compared with 2020, workshop-level, unit-level, and group-level accidents decreased by 13%, 58%, and 43%, respectively. From the perspective of accident types, personal injury and property loss accidents decreased by 19% and 27%, respectively, compared with 2020, and no environmental accident occurred in two years. The accidents have been effectively reduced by formulating measures to solve the two main risk factors of unsafe behaviors, such as weak safety awareness of employees, illegal operation, and frequent equipment damage.

Discussion

The text cluster analyses of workshop-level, unit-level, and group-level accidents are conducted. The result indicates that the accidents of the sample enterprises are mainly personal injury accidents, production accidents, environmental pollution accidents, and quality accidents. The leading causes of personal injury accidents are employees' unsafe behaviors, such as poor safety awareness, non-standard operation, illegal operation, untimely communication, etc. The leading causes of production accidents, environmental pollution accidents, and quality accidents include the unsafe state of materials, such as equipment damage, pipeline leakage, short-circuiting, excessive fluctuation of process parameters, etc. However, some accidents occur under the combined actions of unsafe human behavior and an unsafe material state. For example, employees are suddenly injured by steam pipe network leakage during work, and the operation is not standardized during maintenance, which further leads to excessive fluctuation of process parameters and equipment shut-down.



(a) Accidents at different organizational levels

(b) Different types of accidents

Fig. 11 Comparative analysis of the number of accidents

The risk factors of the enterprise can be effectively identified through accident cluster analysis. Enterprises should strengthen the training of employees' safety knowledge and professional skills, improve the assessment mechanism, and enhance employees' safety awareness and familiarity with the operation process to reduce employees' unsafe behaviors. Given the unsafe state of objects, a more reasonable inspection and maintenance process system should be formulated, and defective equipment and accessories should be replaced in time. In particular, it is necessary to study the maintenance of the frequently faulty equipment identified above, such as the induced draft fan, motor, compressor, distribution room, and substation.

Although workshop-level accidents occur most frequently, the adverse effects are minor, and the hidden risks can be easily eliminated by adopting a series of improvement measures. Due to the small number of group-level accidents, the poor effect of cluster analysis, and the unclear division of accident types and causes, the accident triangle at the organization level can make up for this defect. So, as the number of unit-level accidents is positively related to group-level accidents, the hidden dangers of unit-level accidents can be solved, thus reducing and avoiding group-level accidents.

Conclusion

This article improved the accident triangle and divided chemical accidents into group-level, unit-level, and workshop-level. Based on 484 accident reports of a large chemical enterprise in China, the Spearman correlation coefficient method was used to analyze the rationality of the accident classification. Based on accident classification, TF-IDF and K-means algorithms are used to extract keywords and text clustering analysis is carried out for accidents at all levels. The main conclusions are as follows:

(1) Compared with the traditional accident classification method, the accident triangle proposed in this paper based on the organizational level dramatically reduces the differences between accidents, helps enterprises quickly identify risk factors, and prevents accidents.

(2) There is a significant positive correlation between the unit-level and the group-level accident rate. Enterprises can prevent group-level accidents by reducing the number of unit-level accidents, which solves the problem that it is difficult to find the causes of group-level accidents due to the small number.

(3) This method has achieved significant results in the one-year application process in a large chemical enterprise. Compared with 2020, workshop level, unit level, and group level accounts have been increased by 13%, 58%, and 43%, respectively. In the future, more enterprise data should be selected to validate this method.

Abbreviations

OGP	International Association of Oil and Gas Producers
KDT	Knowledge Discovering in Texts
NLP	Natural Language Processing
TF-IDF	Term Frequency Inverse Document Frequency

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12889-023-17510-w.

Additional file 1. Classification criteria for accidents at different levels.

Additional file 2. Example of accident report.

Additional file 3. Workshop-level accident clustering results.

Acknowledgements

The author thanks the reviewers for their valuable suggestions on improving the quality of the paper. At the same time, I would like to thank the relevant leaders of the safety department of the research enterprise for their support during the author's on-site research.

Authors' contributions

LZ put forward the concept and method of this paper, and completed the preparation of the first draft of the paper. YM proposed the concept of the organization-level accident triangle in this paper. LZM improved the model method of the paper and provided financial support. WXP provided support for the model algorithm in this paper. LTS provided support for the data analysis of this paper. HQR completed the data collection of this paper. SC completed the data collection for this paper.

Funding

The work is funded by the Natural Science Foundation of China (Project Code: 51674193) and the Key R&D Project in Ningxia Hui Autonomous Region (Project Code: 2022BEE02001).

Availability of data and materials

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

Declarations

Ethics approval and consent to participate

This study has passed the ethical review of the Ethics Review Committee of The Chinese Center for Disease Control and Prevention, and the respondents were informed and they consented. The Ethics Review Committee of the China Center for Disease Control and Prevention approved all experimental protocols in this study. All methods of this study were performed following relevant guidelines and regulations. All the participants provided written informed consent, and their participation was voluntary and confidential. Data handling and storage are compatible with this law. All protocols were performed under the Declaration of Helsinki.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 24 March 2023 Accepted: 16 December 2023
Published online: 02 January 2024

References

1. Wang X, Shen Y, Su C. Spatial-temporal evolution and driving factors of carbon emission efficiency of cities in the Yellow River Basin. Energy Rep. 2023;9:1065–70.
2. Lee HE, Kim MG, Yoon SJ, Huh DA, Moon KW. Changes in risk in medium business plating and paint manufacturing plants following the revision of the Korean chemical accident prevention system. Int J Environ Res Public Health. 2021;18:11982.
3. Shi Y, Zhang T, Jiang Y. Digital economy, technological innovation and urban resilience. Sustainability. 2023;15(12):9250.
4. Shan BA, Liu XJ, Gu AW, Zhao RX. The effect of occupational health risk perception on job satisfaction. Int J Environ Res Public Health. 2022;19:2111.
5. Heinrich HW, Stone RW. Industrial accident prevention. Soc Serv Rev. 1931;5(2):323–4.
6. Kines P. Construction workers' falls through roofs: fatal versus serious injuries. J Saf Res. 2002;33(2):195–208.
7. Gnoni MG, Andriulo S, Maggio G, Nardone P. "Lean occupational" safety: an application for a near-miss management system design. Saf Sci. 2013;53:96–104.
8. Wright L, Vander ST. Accident versus near miss causation: a critical review of the literature, an empirical test in the UK railway domain and their implications for other sectors. J Hazrd Mater. 2004;111:105–10.
9. Uth HJ, Wiese N. Central collecting and evaluating of major accidents and near-misses events in the Federal Republic of Germany - results, experiences, perspectives. J Hazrd Mater. 2004;111:139–45.
10. Patterson JM, Shappel SA. Operator error and system, deficiencies: analysis of 508 mining incidents and accidents from Queensland Australia using HFACS. Accident Anal Prev. 2010;42:1379–85.
11. Bourassa D, Gauthier F, Abdul-Nour G. Equipment failures and their contribution to industrial incidents and accidents in the manufacturing industry. Int J Occup Saf Ergo. 2016;2(1):131–41.
12. Liang Cheng SZY, Arnaldo Valdes RM, Gomez Comendador VF, et al. Detection of common causes between air traffic serious and major incidents in applying the convolution operator to Heinrich pyramid theory. Entropy. 2019;21(12):1166.
13. Bird FE, Germain GL. Practical loss control leadership. Loganville, GA: Det Norske Verita; 1996.
14. Okabe M, Ohtani H. Risk estimation for industrial safety in raw materials manufacturing. J Loss Prev Process Ind. 2009;22(2):176–81.
15. Penkey SP, Siddiqui NA. A review on accident pyramid and its empirical interpretation in oil & gas industry (upstream). Int J Sci Res Publ. 2015;5(1):1–3.
16. Manuele FA. Reviewing heinrich dislodging two myths from the practice of safety. Prof Saf. 2011;56(10):52–61.
17. Yorio J, Patrick L, Susan MM. Examining factors that influence the existence of Heinrich's safety triangle using site-Specific H&S data from more than 25,000 establishments. Risk Anal. 2018;38(4):839–52.
18. Rebbitt D. Triangle power: a new view of the great safety triangle. Prof Safety J. 2014;59(09):30–4.
19. Marshall J, Pablo J, Alejandro H, Marcos S. Heinrich's pyramid and occupational safety: a statistical validation methodology. Saf Sci. 2018;101(1):180–9.
20. Shanon HS, Manning DP. Differences between lost-time and non-lost-time industrial accidents. J Occup Accid. 1980;2(4):265–72.
21. Petersen D. Techniques of Safety Management: A Systems Approach. 3rd ed. New York: Aloray Goshen; 1989.
22. Salminen S, Saari J, Saarela KL, Räsänen T. Fatal and non-fatal occupational accidents: identical versus differential causation. Saf Sci. 1992;15:109–18.
23. Hale AR. Conditions of occurrence of major and minor accidents. J Inst Occup Safety Health. 2001;5:7–21.
24. Krause T. Accidents just happen? Industr Safety Hygiene News. 2011;45:1–34.
25. Martin DK, Alison B. Preventing serious injuries & fatalities: study reveals precursors& paradigms. Prof Saf. 2015;60(9):35–43.
26. Saloniemi A, Oksanen H. Accidents and fatal accidents—Some paradoxes. Saf Sci. 1998;29:59–66.
27. Collins RL. Heinrich and beyond. Process Safety Progr. 2011;30(1):2–5.
28. Fu G, Xie XC, Jia QS, Li ZH, Chen P, Ge Y. The development history of accident causation models in the past 100 years: 24Model, a more modern accident causation model. Process Saf Environ. 2020;134:47–82.
29. Cao Y. Research on classification and gradation of offshore oil accident in China. China Safety Sci J. 2022;32(3):18–24.
30. The State Council of the People's Republic of China. "Regulations on Reporting, Investigation and Handling of Production Safety Accidents", Order of the State Council of the People's Republic of China No. 493. 2007. https://www.gov.cn/zwgk/2007-04/19/content_588577.htm.
31. Majumdar A, Manole I, Nalty R. Analysis of Port accidents and calibration of Heinrich's Pyramid. Transport Res Rec. 2022;2676(2):476–89.
32. Yao M, Fang YX, et al. Study on safety behavior planning theory and control strategies for coal chemical workers. Saf Sci. 2020;2020(128): 104726.
33. Wang XP, Zhang C, Deng J, Su C, Gao ZZ. Analysis of factors influencing miners' unsafe behaviors in intelligent mines using a novel hybrid MCDM model. Int J Environ Res Public Health. 2022;19:7368.
34. Dallat C, Salmon PM, Goode N. Risky systems versus risky people: To what extent do risk assessment methods consider the systems approach to accident causation? A review of the literature. Saf Sci. 2019;119:266–79.
35. Liew WT, Adhitya A, Srinivasan R. Sustainability trends in the process industries: a text mining-based analysis. Comput Ind. 2014;65:393–400.
36. Halim SZ, Yu M, Escobar H, Quddus N. Towards a causal model from pipeline incident data analysis. Process Saf Environ Prot. 2020;143:348–60.
37. Steimers A, Schneider M. Sources of risk of AI systems. Int J Environ Res Public Health. 2022;19:3641.
38. Nakata T. Extracting typical incident patterns from text data. Int J Hum Factors Model Simul. 2018;6(2–3):127–39.
39. Victor BG, Perron BE, Sokol RL, et al. Automated identification of domestic violence in written child welfare records: Leveraging text mining and machine learning to enhance social work research and evaluation. J Soc Social Work Res. 2021;12(4):631–55.
40. Feldman R, Dagan I. Knowledge discovery in textual databases (KDT). Proc First Int Conf Knowl Discov Data Mining (KDD-95). 1995;95:112–7 Montreal, Canada, August 20–21.
41. Ananiadou S, Mcnaught J. Text Mining for Biology and Biomedicine; Citeseer: State College. USA: PA; 2006.
42. Vanhala M, Lu C, Peltonen J, Sundqvist S, et al. The usage of large data sets in online consumer behaviour: A bibliometric and computational text-mining–driven analysis of previous research. J Bus Res. 2020;106:46–59.
43. Chen P, Chai J, Zhang L, Wang D. Development and application of a chinese webpage suicide information mining system (sims). J Med Syst. 2014;38(11):88.
44. Fa Z, Li X, Liu Q, et al. Correlation in causality: A progressive study of hierarchical relations within human and organizational factors in coal mine accidents. Int J Environ Res Public Health. 2021;18(9):5020.
45. Li S, You M, Li D, et al. Identifying coal mine safety production risk factors by employing text mining and Bayesian network techniques. Process Saf Environ Prot. 2022;162:1067–81.
46. Na XU, Ling MA, Liu Q, et al. An improved text mining approach to extract safety risk factors from construction accident reports. Saf Sci. 2021;138: 105216.
47. Verma A, Maiti J. Text-document clustering-based cause and effect analysis methodology for steel plant incident data. Int J Inj Contr Saf Promot. 2018;25(4):416–26.
48. Choi YG, Cho KT. Analysis of safety management characteristics using network analysis of CEO messages in the construction industry. Sustainability. 2020;12(14):5771.
49. Lu D, Xu C, Mi C, et al. Establishment of a key hidden danger factor system for electric power personal casualty accidents based on text mining. Information. 2021;12(6):243.
50. Ferreira-Mello R, André M, Pinheiro A, et al. Text mining in education. Wiley Interdiscipl Rev: Data Mining Knowl Discov. 2019;9(6): e1332.

51. Gao L, Wu H. Verb-Based Text Mining of Road Crash Report[C]. TRB 92nd Annual Meeting. 2013.
52. Qiu Z, Liu Q, Li X, et al. Construction and analysis of a coal mine accident causation network based on text mining. Process Saf Environ Prot. 2021;153:320–8.
53. Shahi N, Shahi AK, Phillips R, Shirek G, Lindberg DM, et al. Using deep learning and natural language processing models to detect child physical abuse. J Pediatr Surg. 2021;56(12):2326–32.
54. Raviv G, Fishbain B, Shapira A. Analyzing risk factors in crane-related near-miss and accident reports. Saf Sci. 2016;91:192–205.
55. Brown DE. Text mining the contributors to rail accidents. IEEE Trans Intell Transp Syst. 2016;17:346–55.
56. Song B, Suh Y. Narrative texts-based anomaly detection using accident report documents: the case of chemical process safety. J Loss Prevent Process Ind. 2019;57:47–54.
57. Single JI, Schmidt J, Denecke J. Knowledge acquisition from chemical accident databases using an ontology-based method and natural language processing. Saf Sci. 2020;129: 104747.
58. Ali MXM, Arifin K, Abas A, Ahmad MA, et al. Systematic literature review on indicators use in safety management practices among utility industries. Int J Environ Res Public Health. 2022;19:6198.
59. Zhu YF, Wang DM, Shao ZL, et al. A statistical analysis of coal mine fires and explosions in China. Process Saf Environ. 2019;121:357–66.
60. Cheng CW, Yao HQ, Wu TC. Applying data mining techniques to analyze the causes of major occupational accidents in the petrochemical industry. J Loss Prevent Proc. 2013;26(6):1269–78.
61. Arunraj NS, Mandal S, Maiti J. Modeling uncertainty in risk assessment: an integrated approach with fuzzy set theory and Monte Carlo simulation. Accident Anal Prev. 2013;55:242–55.
62. Uyak V, Ozdemir K, Toroz I. Multiple linear regression modeling of disinfection by-products formation in Istanbul drinking water reservoirs. Sci Total Environ. 2007;378(3):269–80.
63. Boser BE. A training algorithm for optimal margin classifiers. Proc Ann Acm Workshop Comput Learn Theory. 2008;5:144–52.
64. Alpaydin E. Introduction to Machine Learning. 2nd ed. Cambridge, MA: MIT Press; 2011.
65. Jain AK. Data clustering: 50 years beyond k-means. Pattern Recogn Lett. 2010;31(8):651–66.
66. Aggarwal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. ACM SIGMOD Rec. 1993;22(2):207–16.
67. Shirali GA, Noroozi MV, Malehi AS. Predicting the outcome of occupational accidents by CART and CHAID methods at a steel factory in Iran. J Public Health Res. 2018;7(2):1361.
68. Darwiche A. Bayesian networks. Commun ACM. 2010;53:80–90.
69. Singh K, Maiti J, Dhalmahapatra K. Chain of events model for safety management: data analytics approach. Saf Sci. 2019;118:568–82.
70. Wang WF, Xu HJ, Yang WZ. Overview of Chinese word segmentation algorithms. Group technol Prod mod. 2018;35(03):1–8.
71. Zhou WZ, Kan JS. Analysis of food nutrients based on K-means and Apriori algorithm. Sci Techn Eng. 2018;18(17):211–6.
72. Macqueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley symposium on mathematical statistics and probability. Berkeley, USA: University of California Press; 1967. p. 281–97.
73. Peter RJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.