Hazardous Chemical Accident Prevention Based on K-Means Clustering Analysis of Incident Information

FUJIE DENG, WUNAN GU, WENWEN ZENG, ZHENGHUI ZHANG, AND FENG WANG

National Foundation Research Laboratory of Fault Prevention and Hazardous Chemicals Production System, Beijing University of Chemical Technology, Beijing 100029, China

Corresponding author: Feng Wang (wangfeng991@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 51775029, in part by the Chinese Universities Scientific Fund under Grant XK2020-04, and in part by the High Performance Computing Platform of the Beijing University of Chemical Technology (BUCT).

ABSTRACT Hazardous chemicals are inflammable, explosive, and/or toxic and are prone to accidental leakage, fire, and explosion during production, storage, transportation, and transport. It is time-consuming and laborious to study the properties of hazardous chemicals individually for systematic accident prevention because of the wide variety of hazardous chemicals and conditions resulting in accidents. Moreover, accidents have numerous causes, and the relationships among the causative factors are complex. It is a problem that is difficult to accurately identify the effects of correlations among accident factors and determine the laws governing accident occurrence. In this paper, we propose a generic method of hazardous chemical accident prevention based on K-means clustering analysis of incident information to illustrate how to solve the problems. A database of hazardous chemical incidents was constructed, and a K-means clustering algorithm was adopted to classify the incidents. The numbers of occurrences and frequencies of the words in the textual descriptions of the consequences, processes, and causes of hazardous chemical incidents were counted and calculated using a class-based method. For words with a high frequency, risk scenarios were constructed, checklist items of newly revealed dangers were developed, and a system for systematic risk assessment and accident prevention was established. Finally, the information on hazardous material transportation incidents in the Pipeline and Hazardous Materials Safety Administration database of the U.S. Department of Transportation from 2009 to 2018 had been taken as an example to illustrate the method application. The results demonstrate that the proposed method of hazardous chemical accident prevention can be used to improve accident classification. The classification results make it possible to determine the optimal sequence of key targets on which to focus and the requirements for accident prevention and formulate preventive measures. Thus, they provide a technical basis for accident prevention.

INDEX TERMS Hazardous chemicals, accident prevention, incident database, K-means clustering, text analysis.

I. INTRODUCTION

As the quantity of hazardous chemicals in circulation increases, accidental leakage, fires, and explosions are likely to occur during production, storage, transportation, use, and waste disposal, causing casualties, economic losses, and ecological damage [1]–[3]. For example, two explosions occurred in a company’s hazardous chemical warehouse in 2015. The energy of the explosions was equivalent to that of 24 tons of TNT, and the incidents resulted in 165 deaths, 98 injuries, and a direct economic loss of RMB 6.866 billion [4]. The frequent occurrence of hazardous chemical accidents hinders industrial development and decreases social stability. Moreover, people who live near factories have begun to resist the construction and expansion of hazardous chemical production enterprises. Taking preventive measures to avoid the occurrence of hazardous chemical accidents has become an urgent problem for many countries and governments.

Many scholars have investigated the prevention and control of hazardous chemical accidents. The formulation of
protective strategies based on theoretical analyses or simulations of hazardous chemicals or the mechanisms resulting in disasters can reduce the accident rate to some extent. For example, Wang and Zhao [5] established a seamless safety supervision system for the entire process of hazardous chemical handling by analyzing data such as the characteristics and causes of hazardous chemical accidents, thus improving the safety management of hazardous chemicals. Zhang et al. [6] developed a safety audit system to further improve the safety management of hazardous chemicals by analyzing the characteristics of common hazardous chemicals used to manufacture polychlorinated biphenyls and the occurrence mechanisms of common accidents. Nonetheless, many types of hazardous chemicals have different properties. The experimental tests and analysis for each kind of hazardous chemicals, which is time-consuming and labor-intensive, should be conducted to acquire the preventive methods for avoiding accidents. Furthermore, the experimental result is difficult to be applied to the prevention and control of actual accidents.

The incident database can be used to record the information on the incidents referring to hazardous chemicals. The statistical analysis and data mining of the incident information in the database can contribute to taking effective measures to prevent the recurrence of accidents or minimize the severity of accident consequences. Many countries and organizations record incidents in databases that can be used for incident case studies. Examples include the Pipeline and Hazardous Materials Safety Administration (PHMSA) database of the U.S. Department of Transportation, the Process Safety Incident Database (PSID) of the Center for Chemical Process Safety, the Analysis, Research and Information on Accidents (ARIA) database of France, and the Major Accident Reporting System (MARS) of the European Union. The PHMSA database is comprehensive among some available best-known incident databases. The earliest time of the incident information referring to hazardous chemicals in the PHMSA database is 1970. Many scholars have performed statistical analyses of incidents by using the PHMSA database. It is generally because of the ease of obtaining information from PHMSA, the completeness and longer record length of the data, and consistency in the structure of reporting accidents, which contribute to effective, accurate, reliable, and informative data analysis results [7]. For example, the information in incident databases was used to study the mechanism of hazardous chemical accidents, determine the causes of accidents, and qualitatively describe the disaster process [8], [9]. The data was employed to analyze and investigate the causes and consequences of unplanned releases of hazardous chemicals [10]. The annual mileage and accident rates for onshore gas transmission pipelines, and the environmental consequences of historical hazardous liquid pipeline accidents can also be obtained [11]–[13]. However, some incident databases have shortcomings because the data is usually obtained from different sources, and the data is too spotty to compare. It is a challenge to acquire accident laws from a large amount of textual information.

Quantitative analysis of a large number of incident cases combined with automatic clustering, word frequency statistics, text analysis, machine learning, and other computer technologies can be used as an effective method to acquire incident occurrence laws. K-means algorithm is a clustering algorithm based on distance similarity [14]. Due to its advantages of high efficiency and speed, it has been widely used in large-scale data clustering and is an influential clustering algorithm technology. Zhang [15] constructed a case-based reasoning system for emergency treatment of a gas explosion accident based on the statistical analysis results of K-means and KNN. Jain et al. [16] utilized the K-means clustering algorithm to determine the accident-prone states and union territories of India for road accidents in 2012 and applied the decision tree to help discover the causes for accidents. Krishnan et al. [17] used K-means and decision tree algorithms to analyze more than a million data of road accidents from different regions, which had been done using Python and R, to find the common causes of accidents. Chen [18] applied clustering and Pearson correlation analysis to identify the characteristics of 30 typical petrochemical fire accidents and the laws of emergency response and put forward suggestions for prevention and treatment.

This paper proposes a new method that is applying the K-means clustering analysis to acquire hazardous chemical accident laws. The method can contribute to extracting the causes of accidents, acquiring the accident laws from a large amount of textual information, and taking appropriate preventive measures to avoid accidents referring to hazardous chemicals. Firstly, using a K-means clustering algorithm for data mining of incident descriptions, which are text information, will be conducive to the classification of a large number of hazardous chemical incidents. Based on the classification results, the characteristics/label attributes of each category can be obtained, and the accident cause, consequence information, and other disaster-causing laws in each category of data can be extracted from the results of word frequency statistics and probability calculations. Then prevention measures against the identified dangers can be taken to prevent accidents by eliminating the root cause. Finally, data mining of the hazardous material transportation incidents of the PHMSA database from 2009 to 2018 has been taken as an example to illustrate the method application. The components, failure modes, causes of failure, and consequences that need to be checked for each type of accident can be determined according to priority, and preventive measures can be formulated to avoid accidents.

II. OVERVIEW OF HAZARDOUS CHEMICAL ACCIDENT PREVENTION METHOD

The method of hazardous chemical accident prevention based on K-means clustering analysis of incident information has four main components:
A flowchart is shown in Fig. 1.

III. METHOD OF HAZARDOUS CHEMICAL ACCIDENT PREVENTION BASED ON CLUSTER ANALYSIS OF INCIDENT INFORMATION

A. CONSTRUCTION OF HAZARDOUS CHEMICAL INCIDENT DATABASE

The structure and composition of the hazardous chemical incident database were determined according to the accumulated information on hazardous chemical incidents. As shown in Table 1, the incident information is divided into 10 types: (1) basic information on the incident; (2) basic information on the hazardous chemicals; (3) information on the relevant personnel; (4) the status of hazardous chemical equipment; (5) information on packaging or storage; (6) information on transportation; (7) information on the causes; (8) information on the consequences; (9) recommended measures; (10) emergency rescue measures.

The consequences have been divided into several classes, such as spillage, fire, explosion, poisoning, suffocation, and environmental pollution. In this paper, spillage is considered as a type of consequence. The spillage consequence may occur accompanied by other consequences, such as fire, explosion, and environmental pollution, and sometimes not.

B. CLUSTERING ANALYSIS BASED ON INCIDENT DATABASE

1) DATA PREPROCESSING

The hazardous chemical incident database contains mainly numerical and textual data. Before the clustering analysis, the data should first be processed in the following steps.

① Adding an ID column. The ID column can be considered as the unique identification for each accident case.

② Handling with the missing data. Information in some pieces of data may be missing, incorrect, and described by a symbol. To ensure the accuracy of data analysis, it is necessary to fill these data with a value “None.”

③ Converting data type. Standardized and uniform data types are easily recognized by the computer language. The numeric type in the database needs to be converted to a text type.

④ Slicing the words and filtering the stopwords. The field information in each piece of data should be divided into individual words, and then the stopword list can be used to filter out noise.

⑤ Word frequency statistics. Count the number of occurrences of each word in each filtered data, and set a number (key-id value) for each word to build a word index table.

⑥ Determining the key-value value. According to the results of word frequency statistics, the key-value value of each piece of data information can be obtained. The key-value value is composed of the key-id value and the number of occurrences of the word.
2) TEXT ANALYSIS BASED ON LDA MODEL
An unsupervised machine learning technology, latent Dirichlet allocation (LDA), was used to identify hidden topics in hazardous chemical incidents. LDA is a document topic generation model, specifically, a multilayer Bayesian probability model with a three-level document, topic, and word structure [19], [20].

A hazardous chemical incident database can be regarded as a complex document with many topics, where each topic is composed of many words. In the text modeling process using LDA, the total number of documents is denoted as $M$, and the total number of topics is denoted as $K$. In addition, $\alpha$ is the Dirichlet prior parameter of the distributions of multiple topics for each document, and $\beta$ is the Dirichlet prior parameter of the distributions of multiple feature words under each topic [21], [22]. Topics are generated for each document independently; thus, the probability of topic generation in the document library can be calculated using equation (1).

$$p = (Z | \alpha) = \prod_{m=1}^{M} \frac{\Delta(n_m + \alpha)}{\Delta(\alpha)}$$  \hspace{1cm} (1)

For $Z = (z_1, \ldots, z_M)$, $z_m$ is the topic number corresponding to all the words in the $m$-th document; for $n_m = (n^{(1)}_m, \ldots, n^{(K)}_m)$, $n^k_m$ is the number of words in the $k$-th topic in the $m$-th document.

Moreover, words are generated for each of the $K$ topics independently; thus, the probability of word generation in a topic can be calculated using equation (2).

$$p = (W | Z, \beta) = \prod_{k=1}^{K} \frac{\Delta(n_k + \beta)}{\Delta(\beta)}$$  \hspace{1cm} (2)

For $W = (w_1, \ldots, w_M)$, $w_m$ is a word in the $m$-th document; for $n_k = (n^{(1)}_k, \ldots, n^{(T)}_k)$, $n^t_k$ is the number of the $t$-th word produced by the $k$-th topic.

Based on equations (1) and (2), the joint distribution of topics and words, $p(W, Z)$, can be calculated using equation (3).

$$p = (W, Z | \alpha, \beta) = \prod_{k=1}^{K} \frac{\Delta(n_k + \beta)}{\Delta(\beta)} \prod_{m=1}^{M} \frac{\Delta(n_m + \alpha)}{\Delta(\alpha)}$$  \hspace{1cm} (3)

By using the Gibbs sampling algorithm to sample the joint distribution, the probability distribution parameter of the topics in the document ($\hat{\theta}_{mk}$) and the probability distribution parameter of the characteristic words in the topic ($\hat{\phi}_{kt}$) can be obtained using the conventional formula for calculating the Dirichlet distribution. $\hat{\theta}_{mk}$ and $\hat{\phi}_{kt}$ can be calculated using equation (4).

$$\hat{\theta}_{mk} = \frac{n^{(k)}_{m-i} + \alpha_k}{\sum_{k=1}^{K} (n^{(k)}_{m-i} + \alpha_k)}, \quad \hat{\phi}_{kt} = \frac{n^{(t)}_{k-i} + \beta_t}{\sum_{i=1}^{T} (n^{(t)}_{k-i} + \beta_t)}$$  \hspace{1cm} (4)

In summary, the probability of word generation in the document can be calculated using equation (5).

$$p(c_i = k | Z_{-i}, W) \propto \frac{n^{(k)}_{m-i} + \alpha_k}{\sum_{k=1}^{K} (n^{(k)}_{m-i} + \alpha_k)}, \quad \frac{n^{(t)}_{k-i} + \beta_t}{\sum_{i=1}^{T} (n^{(t)}_{k-i} + \beta_t)}$$  \hspace{1cm} (5)

where $i(m, n)$ is a two-dimensional subscript, $c_i = k$ indicates that the topic of the $n$-th word in the $m$-th document is $k$, and $-i$ indicates that the word represented by subscript $i$ is removed.

The number of topics in the incident process description in the hazardous chemical incident database is set to 50. The LDA algorithm can be used to calculate the frequencies with which the incident corresponding to each topic appear.

3) K-MEANS CLUSTERING ANALYSIS
The K-means clustering algorithm is widely used for text analysis [23]. It can be applied to divide the incident text information into several categories according to specific rules so that each type of accident will be given some specific characteristics/label attributes. The characteristics/label attributes will be beneficial to statistical analysis of accidents, and some new laws can be excavated from a large number of data, which provides a reference for the prevention of accidents.

In this study, K-means clustering was adopted to classify the information from the incident process descriptions in the hazardous chemical incident database, as follows.

---

**TABLE 1. Structure and field names of hazardous chemical incident database.**

| ID | Information classification | Type of information |
|----|---------------------------|---------------------|
| 1  | Basic information on the incident | Incident name, date, time, location, weather conditions, and incident process description. |
| 2  | Basic information on the hazardous chemical | Hazardous chemical name, hazard overview, physicochemical properties, and storage capacity. |
| 3  | Information on the relevant personnel | Carrier name, carrier city, carrier state, carrier HAZMAT Reg ID, and shipper name. |
| 4  | Status of hazardous chemical equipment | Names and technical specifications of equipment. |
| 5  | Information on packaging or storage | Packaging type, packaging material, and packaging/storing volume (net weight). |
| 6  | Information on transportation | Transportation phase, facility, and regulations. |
| 7  | Cause | Failed component, failure mode, and cause of failure. |
| 8  | Consequence | Spillage, fire, explosion, poisoning, suffocation, environmental pollution (water, air, soil, and biology), casualties (fatalities and injuries), and economic damages. |
| 9  | Recommended measures | Production measures, management measures, storage measures, and transportation measures. |
| 10 | Emergency rescue measures | Emergency plan, relevant laws and regulations, and accident handling. |
The probability values of the incident corresponding to the 50 topics calculated by the LDA model are combined with the ID column and the key-value value of the data; thus, each data vector contains 52 columns of attributes.

Suppose that the incident process description information is divided into 30 categories, where the self-similarity of each category should be maximized. The self-similarity is calculated using the average value of the data in each class.

One data item is randomly selected from each class, for a total of 30 data items, each of which initially represents the average or center of its class.

Each of the remaining data items is assigned to the nearest class according to its distance from the center of the class, and the average value of each class is recalculated.

The number of iterations until the criterion function converges is set so as to minimize the sum of the mean square error within each class. The K-means clustering algorithm stops when the center point of each class no longer changes.

If X and Y are two data items, the expressions are
\[
X = (x_1, x_2, \ldots, x_n) \quad \text{and} \quad Y = (y_1, y_2, \ldots, y_n),
\]
each has 52 characteristic attributes (i.e., \( n = 52 \)). The Euclidean distance between \( X \) and \( Y \) can be obtained using equation (6).

\[
\text{dist} (X, Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2} \tag{6}
\]

A new class center point is obtained by calculating the average value of each assigned point in the current class. The class boundary is adjusted according to the relocated center. The process of updating and distributing is repeated multiple times; finally, the classification data for the 30 classes is obtained. The K-means clustering algorithm can be implemented by some software such as Python, R, or the Alibaba Cloud’s Platform of Artificial Intelligence (PAI).

C. WORD FREQUENCY STATISTICS AND ANALYSIS OF CLASSIFIED ACCIDENT INFORMATION

To determine the correlations among the consequences, processes, and causes of hazardous chemical incidents and identify the key factors for accident prevention, it is necessary to obtain the word frequency statistics and analyze the classification results. These results can indicate the importance of each word for describing hazardous chemical accidents.

(1) Statistical analysis of the consequences of hazardous chemical incidents

There are six main types of consequences of incidents: spillage, fire, explosion, poisoning, suffocation, environmental pollution. Some accident consequences may include only one class, and some may include more than two classes. The number of different consequences of each type of accident in the database is counted, and the frequency of each consequence among all the consequences for that accident type is calculated. According to the analysis and comparison of the probability of each consequence, the consequences with high probability can be taken as key targets for accident prevention, and careful inspection can be performed during risk assessment to prevent and control each type of accident.

(2) Statistical analysis of the process description information for hazardous chemical incidents

The word frequency statistics are obtained for the 30 types of data resulting from K-means clustering, and the words are sorted. Several words with the highest word frequency can be selected as key factors for all 30 types of data. Moreover, for the selected words, the number of occurrences and frequency of occurrence in the 30 types are counted and calculated. On the basis of the probability of each word, the targets requiring attention and detailed inspection for each type of accident can be ordered according to priority, and effective measures can be taken to prevent accidents during processing.

(3) Statistical analysis of the causes of hazardous chemical incidents

The number and frequency of the cause data items are counted and calculated by category. The most common words and those with the highest frequency among the 30 accident types are selected. The key factors for the 30 types are obtained by combining and counting these words. By calculating the frequency of occurrence of these factors in each category, the causal factors requiring attention and detailed inspection can be ordered according to priority. Then preventive measures against the identified dangers can be taken to prevent accidents by eliminating the root cause.

D. DEVELOPMENT OF ACCIDENT PREVENTION INFORMATION SYSTEM

The statistical analysis results reveal the key factors of incident consequences, processes, and causes for all the categories. By matching and combining these factors, and constructing risk scenarios, relevant checklist items can be proposed according to the identified hazards. An accident prevention information system based on the results of the summary and classification of the checklist items can be developed. The system can be applied to accurately identify the factors that are likely to cause hazardous chemical accidents and thus to provide technical guidance for the prevention and control of hazardous chemical accidents.

IV. APPLICATION

Using the above models and method, we constructed an incident database based on PHMSA data from 2009 to 2018 and quantified and classified the case information. An accident prevention information checklist based on the classification results was compiled, and an accident prevention information system for hazardous chemical transportation was developed. The checklist and system can be used to improve safety inspections on the basis of the risks associated with transportation of hazardous chemicals and the identified dangers.
A. OVERVIEW OF INFORMATION IN THE HAZARDOUS CHEMICAL TRANSPORTATION INCIDENT DATABASE

1) CONSTRUCTION AND STRUCTURE OF INCIDENT DATABASE

PHMSA began to record hazardous chemical transportation incident information in 1970, and more than 600,000 items have been stored to date. These reports contain information on commodity properties, hazardous class, packing group and type, incidents’ causes, and resulting consequences that are helpful for safety research agencies, government departments, and industries personnel who conduct the inspection, planning, and risk assessment activities. Clustering analysis was applied to the incident information recorded in the last 10 years (2009–2018). The number of hazardous chemical transportation incidents per year during this period is shown in Fig. 2.

FIGURE 2. Number of hazardous chemical transportation incidents per year from 2009 to 2018.

A hazardous chemical transportation incident database was constructed using the structure shown in Table 1. The information in the “cause” and “consequence” was refined based on the PHMSA data. The text information corresponding to “description of events,” “cause,” and “consequence” in the database was quantified and analyzed. The field names corresponding to the causes and consequences are shown in Table 2.

TABLE 2. Field names corresponding to causes and consequences.

| Information classification | Type of information |
|----------------------------|---------------------|
| Cause                      | What failed code, what failed description, how failed code, how failed description, failure cause code, failure cause description, etc. |
|                            | Spillage, fire, explosion, material entered waterway or storm sewer, gas dispersion, environmental damage, economic loss, hazmat fatalities, hazmat hosp injuries, hazmat nonhosp injuries, etc. |
| Consequence                |                      |

2) CLASSIFICATION OF CAUSES OF HAZARDOUS CHEMICAL TRANSPORTATION INCIDENTS

Depending on the cause class in the PHMSA database, the causes of the incidents in the hazardous chemical transportation incident database are divided into three categories: the failed component, failure mode, and cause of failure. Moreover, according to the descriptions in the PHMSA database, the specific items in each category are subdivided into several categories.

① Failed component

The components causing an accident are divided into 61 categories, including valves, cylinders, gaskets, metering devices, alarm devices, inlets and outlets, pipelines, bolts, flanges, and other accessories. A partial list of the failed components and codes is shown in Table 3.

TABLE 3. Partial list of failed components and codes.

| Code   | What failed code | What failed          |
|--------|------------------|----------------------|
| 101    | Air inlet        | Cylinder neck or shoulder |
| 102    | Auxiliary valve  | Cylinder sidewall - near base |
| 103    | Basic material   | Cylinder sidewall - other |
| 104    | Body             | Cylinder valve        |
| 105    | Bolts or nuts    | Discharge valve or coupling |
| 106    | Bottom outlet valve | Excess flow valve |
| 107    | Check valve      | Fill hole            |
| 108    | Chime            | Flange               |
| 109    | Closure (e.g., cap, top, or plug) | Frangible disc |
| 110    | Cover            | Fusible pressure relief device or element |

② Failure mode

The failure modes leading to accidents are classified into 13 categories, including abrasion, bending, rupture, crushing, and leakage. Table 4 lists the failure modes and corresponding codes.

TABLE 4. Failure modes and codes.

| Code | How failed | Code | How failed |
|------|------------|------|------------|
| 301  | Abraded    | 308  | Leaked     |
| 302  | Bent       | 309  | Punctured  |
| 303  | Burst or ruptured | 310  | Ripped or torn |
| 304  | Cracked    | 311  | Structural |
| 305  | Crushed    | 312  | Torn off or damaged |
| 306  | Failed to operate | 313  | Vented     |
| 307  | Gouged or cut |      |            |

③ Cause of failure

The causes of accidents or component failures are classified into 38 categories, including human error, commodity self-ignition, commodity polymerization, improper preparation for transportation, and rollover accident. A partial list of the causes of failure and codes is shown in Table 5.

B. CLUSTER ANALYSIS AND WORD FREQUENCY STATISTICAL ANALYSIS BASED ON HAZARDOUS CHEMICAL TRANSPORTATION INCIDENT DATABASE

The data on the descriptions of events in the hazardous chemical transportation incident database were preprocessed, including handling missing data, converting data types, and
TABLE 5. Partial list of causes of failure and codes.

| Code | Failure cause               | Code | Failure cause               |
|------|----------------------------|------|----------------------------|
| 501  | Abrasion                   | 511  | Dropped                    |
| 502  | Broken component or device | 512  | Fire, temperature, or heat |
| 503  | Commodity self-ignition    | 513  | Forklift accident          |
| 504  | Commodity polymerization   | 514  | Freezing                   |
| 505  | Conveyor or material handling equipment mishap | 515  | Human error               |
| 506  | Corrosion - exterior       | 516  | Impact with sharp or protruding object (e.g., nails) |
| 507  | Corrosion - interior       | 517  | Improper preparation for transportation |
| 508  | Defective component or device | 518  | Inadequate accident damage protection |
| 509  | Derailment                 | 519  | Inadequate blocking and bracing |
| 510  | Deterioration or aging     | 520  | Inadequate maintenance     |

filtering stopwords. For example, there were 17,384 incident cases recorded in 2017, while 344 cases did not have a description of events information. Therefore, we handle the missing data with a blank value of “None” before clustering to ensure that each piece of data participates in the clustering calculation.

LDA model and K-means clustering algorithms were applied to the processed data to classify the incidents into 30 categories, and the classified data tables containing all the field information of the hazardous chemical incident database can be obtained. If we want to get a more accurate disaster-causing law, the LDA model and K-means clustering algorithm can be used to conduct classification and data mining for each type of data again.

Based on the classified data tables, the number and frequency of the words in the text descriptions of the consequences, process descriptions, and causes of the incidents in each class were counted and calculated. For the words that appear with high frequency, the characteristics/label attributes of each type of accident can be determined, and a specific label can be defined to describe the accident characteristically. According to the defined label, the accident information of the category corresponding to the label can be obtained, and the most probable accident factors can be further identified, which can facilitate the subsequent identification of key targets for inspection in light of the identified dangers, and thus reduce the likelihood of transportation accidents.

In the following analysis, the consequences of incidents are taken as the starting point, and the category with the most consequences is chosen as an example to analyze the correlations among the consequences, processes, and causes of the incident.

1) STATISTICAL ANALYSIS OF THE CONSEQUENCES OF HAZARDOUS CHEMICAL TRANSPORTATION INCIDENTS

The number and frequency of different consequences of each type of incident were counted and calculated to analyze the consequences of hazardous chemical transportation incidents. Six types of consequences were counted: spillage, fire, explosion, material entered waterway or storm sewer (hereinafter referred to as water sewer), gas dispersion, and environmental damage.

![FIGURE 3. Number of occurrences of six types of consequences.](image)

Fig. 3 shows the number of occurrences of each type of consequence. The three most common consequences are spillage, gas dispersion, and fire, which can be used as the key factors of concern.

Fig. 4 shows the data curves for the occurrence of spillage for the 30 incident categories. The bar chart indicates the total number of consequences for each type of incident. The orange line represents the number of occurrences of spillage in each type of incident, and the green line graph represents the relative frequency of spillage among all the consequences in each category. Class 23 exhibits the most consequences and occurrences of spillage.

Fig. 5 shows the data curves for other types of consequences in the 30 categories. A comparison of Fig. 4 and Fig. 5 reveals that spillage accounts for a much higher proportion of the consequences than the other consequences in each category. According to the probability of each consequence in each category, enterprises can prioritize the prevention of the most likely consequences. Thus, the corresponding preventive measures can be formulated to avoid or reduce the loss caused by accidents. For Class 23, the key consequences are spillage and gas dispersion.
2) STATISTICAL ANALYSIS OF THE PROCESS DESCRIPTION INFORMATION FOR HAZARDOUS CHEMICAL TRANSPORTATION INCIDENTS

To analyze the incident process descriptions, the 10 words occurring most frequently in all the categories and the number of occurrences of each one in the 30 categories are determined. The statistical analysis of the 10 words and their assignment as key factors of concern in all the categories can guide inspections for the identified dangers and the design of safety checklist items.

Fig. 6 shows the frequency of occurrence of the 10 words that appear most frequently in the 30 categories, which are package, container, product, drum, freight, car, release, trailer, driver, and leaking. Among them, package accounts for 26.64%, which is the highest word frequency in all the categories.

Fig. 7 shows the frequency of occurrence of some of the words with the highest frequency in the 30 categories. The line chart shows the result for package, and points represent the results for other words. The frequency of a word can be regarded as the probability of its occurrence in each category. By comparing these probabilities, the sequence in which process factors need to be checked to prevent each type of accident can be determined. According to this method, a line chart can be made for each word depending on the purpose of the research.

A comprehensive comparison of the probability of the 10 words with the highest word frequency in Class 23 reveals that the optimal order of inspection is package > container > product > freight > driver > drum > release > car.

3) STATISTICAL ANALYSIS OF THE CAUSES OF HAZARDOUS CHEMICAL TRANSPORTATION INCIDENTS

The two words with the highest frequency in the failed component, failure mode, and cause of failure information for the 30 categories were identified to analyze the causes of hazardous chemical transportation incidents. By combining and counting the occurrences of these 60 words, we obtained the key factors of concern for the failed component, failure mode, and cause of failure. The frequency of occurrence of these key factors in each category was calculated; this information can guide analysis by identifying potential risks and highlighting the factors that are likely to cause accidents.

The analysis results can be applied to guide accident investigation and analysis to ensure the completeness of the
analysis and avoid the omission of potential risk factors, such as failed components, failure modes, and causes of failure.

1. Statistical analysis of failed component

By combining the statistics of the two factors with the highest word frequency in the failed component information for the 30 categories, we obtained a total of nine failure components and took them as the key factors of concern. Fig. 8 shows the frequency of occurrence of these factors in the 60 selected data items. The components most vulnerable to failure in all categories are closures (e.g., caps, tops, or plugs), the body, and the basic material.

Fig. 9 shows the frequency of the three most common failure modes in the 30 categories. The line chart shows the frequency of leaking, and the points represent the frequencies of other failure modes. These frequencies can be regarded as the probability of occurrence in each category. By comparing the probabilities of the failure modes, the sequence in which failure modes should be checked to prevent each type of accident can be determined, and preventive measures can be taken at an early stage of the accident.

A comprehensive comparison of the probability of the seven failure modes in Class 23 reveals that the optimal order of inspection is leaking > punctures > cracks > crushing > failure to operate > bursting or rupture > venting.

2. Statistical analysis of failure mode

By combining the statistics of the two factors with the highest word frequency in the failure mode information for the 30 categories, we obtained a total of seven failure modes and took them as the key factors of concern. Fig. 10 shows the frequency of occurrence of these factors in the 60 selected data items. The most common failure modes in all categories are leaking, puncture, and crushing.

Fig. 11 shows the frequency of the three most common failure modes in the 30 categories. The line chart shows the frequency of leaking, and the points represent the frequencies of other failure modes. These frequencies can be regarded as the probability of occurrence in each category. By comparing the probabilities of the failure modes, the sequence in which failure modes should be checked to prevent each type of accident can be determined, and preventive measures can be taken at an early stage of the accident.

A comprehensive comparison of the probability of the seven failure modes in Class 23 reveals that the optimal order of inspection is leaking > punctures > cracks > crushing > failure to operate > bursting or rupture > venting.

3. Statistical analysis of cause of failure

By combining the statistics of the two factors with the highest word frequency in the cause of failure in the 30 categories, we obtained a total of 13 causes of failure and took them as
the key factors of concern. Fig. 12 shows the frequency of occurrence of these factors in the 60 selected data items. The most common causes of failure in all the categories are loose closure (component or device), forklift accident, and human error.

Fig. 13 shows the frequencies of the three most common causes of failure in the 30 categories. The line chart shows the frequency of “loose closure (component or device),” and the points represent the frequencies of other causes of failure. The frequencies of the causes of failure can be regarded as the probability of occurrence in each category. A comparison of the probabilities of each cause reveals the optimal sequence in which to check the causes to prevent each type of accident; consequently, preventive measures can be taken to avoid accidents.

According to a comprehensive comparison of the probability of the 13 causes of failure in Class 23, the optimal order of inspection is loose closure (component or device) > forklift accident > dropping > defective component or device > improper preparation for transportation > human error > impact with sharp or protruding object (e.g., nails) > inadequate preparation for transportation > inadequate blocking and bracing > too much weight on package > deterioration or aging > rollover accident > vehicular crash or accident damage.

### 4) ANALYSIS OF CORRELATIONS AMONG ACCIDENT FACTORS

The comprehensive analysis of the cause, process, and consequences of each type of incident reveals a causal linkage between the factors contributing to the accident. Consequently, an accident analysis diagram can be drawn for each type of incident. We take Class 23, which has the largest number of consequences, as an example to illustrate the accident analysis. The diagram in Fig. 14 shows the two items with the highest frequency in the incident process description; failed component, failure mode, and cause of failure information; and consequences, along with the corresponding probabilities.

For the incidents in Class 23, spillage and gas dispersion are the most common consequences; package and container are the two most frequently appearing words in the incident process descriptions. In addition, closures (e.g., caps, tops, or plugs) and bodies are the most vulnerable components, leaking and puncture are the most common failure modes of components, and loose closures (components or devices) and forklift accidents are the most common causes of component failure. Risk scenarios can be constructed on the basis of the statistical analysis of the consequences, processes, and causes of incidents. During hazardous chemical transportation, loose closures caused storage facilities or containers to leak. Forklift accidents caused packages or bodies to be punctured, leading to cracks in storage facilities and material leakage. Therefore, the most common consequence of this type of incident is spillage. The analysis results can verify the causal chain leading to the accident; that is, the occurrence, development, and consequences of the accident have a clear causal relationship.

The direct causes of accidents are unsafe states of objects and unsafe human behavior. The results of the analysis can guide enterprises to focus on the equipment components and causes of failure that may lead to accidents. Consequently, preventive measures can be formulated to avoid unsafe states of objects and unsafe human behavior during the transportation of hazardous chemicals and to eliminate the underlying dangers. If the enterprise can eliminate these dangers in any step in the hazardous chemical transportation process, it can destroy the causal chain of the accident and halt or control the accident process.

### C. DEVELOPMENT OF ACCIDENT PREVENTION INFORMATION SYSTEM FOR HAZARDOUS CHEMICAL TRANSPORTATION

#### 1) IDENTIFY THE KEY FACTORS

The incident process descriptions, causes (failed component, failure mode, and cause of failure), and consequences of incidents in a hazardous chemical transportation incident database were statistically analyzed. Table 6 shows the key factors for preventing hazardous chemical transportation incidents.

By reasonably matching the above factors according to the form “incident process description/failed component +
TABLE 6. Key factors for preventing hazardous chemical transportation incidents.

| Item               | Key factors                                                                                                                                 |
|--------------------|---------------------------------------------------------------------------------------------------------------------------------------------|
| Description of events | Package, container, product, drum, freight, car, release, trailer, driver, and leaking.                                                       |
| Failed component   | Closure (e.g., cap, top, or plug), body, basic material, inner packaging, manway or dome cover, hose, tank shell, liquid valve, and bottom outlet valve. |
| Failure mode       | Leaked, punctured, crushed, cracked, failed to operate, burst or ruptured, and vented.                                                         |
| Cause of failure   | Loose closure (component or device), forklift accident, human error, dropped, inadequate blocking and bracing, improper preparation for transportation, impact with sharp or protruding object (e.g., nails), defective component or device, inadequate preparation for transportation, deterioration or aging, vehicular crash or accident damage, rollover accident, and too much weight on package. |
| Consequence        | Spillage, gas dispersion, and fire.                                                                                                         |

FIGURE 14. Accident analysis for Class 23.

failure mode + cause of failure + consequence,” risk scenarios can be constructed, and the corresponding security inspection items can be identified. An accident prevention information checklist based on the inspection items can be established to evaluate the risk of the identified dangers systematically, and safety management personnel can propose targeted prevention and control measures to avoid accidents.

2) DETERMINE THE HIERARCHICAL STRUCTURE OF THE ACCIDENT PREVENTION INFORMATION CHECKLIST
The inspection items were summarized and categorized in this study according to the key factors and proposed safety inspection items, and the hierarchical structure of the accident prevention information checklist was determined, as shown in Table 7.

3) FUNCTION AND APPLICATION
Using the hierarchical structure of the accident prevention information checklist, we stored the safety inspection items in a database, and we designed and developed an accident prevention information system for hazardous chemical transportation. The system interface is shown in Fig. 15.

The accident prevention information system has multiple functional options that can be used to select the analysis factor category. By clicking on the subcategories, users can further
refine the selection to facilitate the addition of information to the accident prevention information checklist. In addition, the checklist can be obtained more rapidly. The system can add, save, print, and modify the information on the checklist.

If the logistics company undertakes a new task to deliver hazardous chemicals to some places, the carriers or supercargos can identify the potential risk that may lead to incidents according to the analysis results. As a result, when using the accident prevention information system, they can perform inspections in the optimal order to prioritize the key factors that are likely to cause hazardous chemical transportation accidents. When these inspections are combined with feedback on relevant problems in the checklist report, effective measures can be taken to reduce or avoid the risk of hazardous chemical transportation.

V. CONCLUSION
We presented a generic method of hazardous chemical accident prevention based on K-means clustering analysis of incident information. In this paper, a hazardous chemical incident database was constructed. A K-means clustering algorithm was adopted to classify hazardous chemical accidents, which will help field experts or safety researchers subdivide a large number of accidents and provide references for accident classification. Besides, the K-means clustering method was utilized to conduct a preliminary analysis before the statistical analysis of the incident cases, which can contribute to extracting the causes of accidents and acquiring the laws from a large amount of textual information. Then the causes, consequences, and processes of each type of accident were counted and analyzed. By calculation and analysis, the factors prone to failure were selected, accident risk scenarios were constructed to identify key monitoring targets for safety managers, and safety inspection items against the identified dangers were established. Based on these inspection items, an accident prevention information system for hazardous chemicals was developed to formulate more effective prevention and control measures. Moreover, a comprehensive analysis of the correlations among the causes, processes, and consequences of each type of accident clarified the causal chains of accidents. Finally, the PHMSA database had been taken as an example to illustrate the method application.

The proposed method can be applied to prevent accidents referring to hazardous chemicals. With the aid of the accident prevention information system developed for hazardous chemicals, the safety and reliability of the production, usage, transportation, and storage of hazardous chemicals will be improved.

REFERENCES
[1] L. Guo, C. Jin, and G. Liu, “Design of monitoring system for hazardous chemicals logistics based on Internet of Things,” in Proc. 3rd IEEE Int. Conf. Comput. Commun. (ICCC), Chengdu, China, Dec. 2017, pp. 1168–1173.
[2] B. Wang, C. Wu, G. Reniers, L. Huang, L. Kang, and L. Zhang, “The future of hazardous chemical safety in China: Opportunities, problems, challenges and tasks,” Sci. Total Environ., vol. 643, pp. 1–11, Dec. 2018.
[3] C. Chen and G. Reniers, “Chemical industry in China: The current status, safety problems, and pathways for future sustainable development,” Saf. Sci., vol. 128, Aug. 2020, Art. no. 10471.

[4] Z. Lu, G. Fu, and Z. Z. Xue, “Research on behavioural causes of a fire and explosion accident of ‘8-12’ in Tianjin port,” J. Catastrophol., vol. 32, no. 1, pp. 205–211, Jan. 2017.

[5] X. L. Wang and L. J. Zhao, “How to improve the safety management of hazardous chemicals in China-based on the big data analysis of hazardous Accidents,” Explor. Free Views, vol. 2, pp. 73–77, Feb. 2016.

[6] Z. D. Zhang and D. Zhang, “Accident analysis and preventive safety risk management of hazardous chemicals in PCB enterprises,” Printed Circuit Inf., vol. 27, no. 3, pp. 51–59, Mar. 2019.

[7] C. Belvederesi, M. S. Thompson, and P. E. Komers, “Canada’s federal database is inadequate for the assessment of environmental consequences of oil and gas pipeline failures,” Environ. Rev., vol. 25, no. 4, pp. 415–422, Dec. 2017.

[8] M. Lecue and R. M. Durbra, “Accidents in European ports involving chemical substances: Characteristics and trends,” Saf. Sci., vol. 115, pp. 278–284, Jun. 2019.

[9] R. Moura, M. Beer, E. Patelli, J. Lewis, and F. Knoll, “Learning from major accidents to improve system design,” Saf. Sci., vol. 84, pp. 37–45, Apr. 2016.

[10] C. E. Restrepo, J. S. Simonoff, and R. Zimmerman, “Causes, cost consequences, and risk implications of accidents in US hazardous liquid pipeline infrastructure,” Int. J. Crit. Infrastruct. Protection, vol. 2, nos. 1–2, pp. 38–50, May 2009.

[11] C. Lam, “Statistical analyses of historical pipeline incident data with application to the risk assessment of onshore natural gas transmission pipelines,” M.S. thesis, Univ. Western Ontario, London, ON, Canada, 2015.

[12] C. Lam and W. Zhou, “Statistical analyses of incidents on onshore gas transmission pipelines based on PHMSA database,” Int. J. Pressure Vessels Piping, vol. 145, pp. 29–40, Sep. 2016.

[13] C. Belvederesi, M. S. Thompson, and P. E. Komers, “Statistical analysis of environmental consequences of hazardous liquid pipeline accidents,” Heliyon, vol. 4, no. 11, Nov. 2018, Art. no. e00901.

[14] D. M. Tang, “Study on clustering algorithm and its applications,” Ph.D. dissertation, Univ. Electron. Sci. Technol. China, Chengdu, China, 2010.

[15] L. Y. Zhang, “Study and application of emergency decision on case-based reasoning of gas in coal mine,” M.S. thesis, China Univ. Mining Technol., Xuzhou, China, 2014.

[16] A. Jain, G. Ahuja, Anuranjana, and D. Mehrotra, “Data mining approach to analyse the road accidents in India,” in Proc. 5th Int. Conf. Rel., Infocom Technol. Optim. (Trends Future Directions) (iCRITO), Noida, India, Sep. 2016, pp. 175–179.

[17] P. V. Krishnan, V. Chandra Sheel, M. V. S. Viswanadh, C. Shetty, and S. Seema, “Data analysis of road traffic accidents to minimize the rate of accidents,” in Proc. 3rd Int. Conf. Comput. Syst. Inf. Technol. Sustain. Solutions (CSITSS), Bengaluru, India, Dec. 2018, pp. 247–253.

[18] Z. N. Chen, L. Z. Wu, and Q. L. Wang, “Analysis of 30 typical petrochemical fire accidents based on cluster model and correlation coefficient,” China Public Secur. (Acad. Ed.), vol. 4, no. 1, pp. 205–211, Jan. 2019.

[19] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” J. Mach. Learn. Res., vol. 3, pp. 993–1022, Mar. 2003.

[20] J. L. Dong, “Document clustering method based on LDA topic model,” M.S. thesis, Central China Normal Univ., Wuhan, China, 2012.

[21] X. Li, J. Q. Gao, and J. Y. Yang, “Image scene classification based on PLDA in Hadoop,” J. NanJing Jiaotong Univ., vol. 34, no. 6, pp. 81–86, Dec. 2015.

[22] P. Yang, Y. Yao, and H. Zhou, “Leveraging global and local topic popularities for LDA-based document clustering,” IEEE Access, vol. 8, pp. 24734–24745, Jan. 2020.

[23] M. J. Li, M. K. Ng, Y.-M. Cheung, and J. Z. Huang, “Agglomerative fuzzy K-means clustering algorithm with selection of number of clusters,” IEEE Trans. Knowl. Data Eng., vol. 20, no. 11, pp. 1519–1534, Nov. 2008.

FUJIE DENG received the B.S. degree in safety engineering from Southwest Petroleum University, in 2018. She is currently a Graduate Student of safety science and engineering with the College of Mechanical and Electrical Engineering, Beijing University of Chemical Technology. Her research interests include chemical process safety and early warning.

WUNAN GU is currently pursuing the bachelor’s degree in safety engineering with the College of Mechanical and Electrical Engineering, Beijing University of Chemical Technology. She was admitted to the Beijing University of Chemical Technology, in 2016. Her research interests include chemical process safety and early warning.

WENWEN ZENG received the B.S. degree in safety engineering from the Beijing University of Chemical Technology, in 2017. She is currently a Graduate Student of mechanic and electronic engineering with the Beijing University of Chemical Technology. Her research interests include chemical engineering safety and data mining.

ZHENGHUI ZHANG received the B.S. degree in process engineering and control engineering from the Beijing University of Chemical Technology, in 2018. He is currently a Graduate Student of mechanic and electronic engineering with the Beijing University of Chemical Technology. His research interests include machine design and equipment fault diagnosis.

FENG WANG received the Ph.D. degree in chemical process machinery from the Beijing University of Chemical Technology, in 2009. He is currently an Associate Professor of mechanic and electronic engineering with the Beijing University of Chemical Technology. His research interests include equipment fault diagnosis, chemical process safety and early warning, and intelligent diagnosis technology based on multi-parameter data fusion.