An Improved Text Mining Approach to Extract Safety Risk Factors from Construction Accident Reports

XU, N [a,b]; **MA, L** [c*]; Liu, Q [d]; WANG, L [e]; Deng, Y [f]

[a] School of Mechanics & Civil Engineering, China University of Mining and Technology, Xuzhou, China
[b] State Key Laboratory for Geomechanics and Deep Underground Engineering, Xuzhou, China
[c] School of Bartlett Construction and Project Management, University College London, London, UK
[d] School of Civil Engineering, Xuzhou University of Technology, Xuzhou, China
[e] School of Mechanics & Civil Engineering, China University of Mining and Technology, Xuzhou, China
[f] School of Mechanics & Civil Engineering, China University of Mining and Technology, Xuzhou, China

Abstract

Workplace accidents in construction commonly cause fatal injury and fatality, resulting in economic loss and negative social impact. Analyzing accident description reports helps identify typical construction safety risk factors, which then becomes part of the domain knowledge to guide safety management in the future. Currently, such practice relies on domain experts' judgment, which is subjective and time-consuming. This paper developed an improved approach to identify safety risk factors from a volume of construction accident reports using text mining (TM) technology. A TM framework was devised, and a workflow for building a tailored domain lexicon was established. To reduce the impact of report length, information entropy weighted term frequency ($TF-H$) was proposed for term-importance evaluation, and an accumulative $TF-H$ was proposed for threshold division. A case study of metro construction projects in China was conducted. A list of 37 safety risk factors was extracted from 221 metro construction accident reports. The result shows that the proposed $TF-H$ approach performs well to extract important factors from accident reports, solving the impact of different report lengths. Additionally, the obtained risk factors depict a portrait of critical causes contributing most to metro construction accidents in China. Decision-makers and safety experts can use these factors and their importance degree while identifying safety factors for the project to be constructed.

Keywords: construction safety; automatic risk identification; workplace accident; text mining

1. Introduction

Project risk is defined as an uncertain event or condition that, if it occurs, has a positive or a negative effect on at least one project objective (PMI 2017). In the context of occupational health and safety, risk is defined as the factor that might cause accidents in a work environment (Karasan et al. 2018). Safety risk management identifies and controls the associated risks that may lead to accidents (Dallat et al. 2019), thus, benefits to minimize the possible losses and damages resulting from work-related, worksite-related, and worker-related activities (Gul and Ak 2018). As the first step of safety risk management, the identification of safety risk factors is vital for assessing risk status and planning mitigation actions (Gul 2018). In the construction industry, safety risk identification frequently relies on professional estimates to determine the possible factors. Professionals use their learning-from-past experience, an essential source of domain knowledge, to identify safety risks.

Experience, as tacit knowledge, embedded in the human mind, is difficult and costly to obtain. Researchers have used tools, such as brainstorming, Delphi method, questionnaires, interview, cause-and-effect analysis, literature study and their combination (Qazi et al. 2016; Soliman 2018; Tembo-Silungwe and Khatleli 2018) to encapsulate the domain knowledge. These traditional data collection methods usually need a certain amount of experienced experts and consume extensive time and cost. While collecting data from a small number of experts may lead to an incomplete and biased risk checklist.

Text information, as explicit knowledge, codified and digitized in documents and reports, is easy to be shared (Nonaka 2008). In the construction industry, accident reports are used to record the causes, consequences, and the whole process of accidents. Hundreds of accident reports make a valuable knowledge database. Researchers have been using conventional descriptive statistics to summarize key safety risk factors from those reports (Rivas et al. 2011). However, as the information hidden in the reports is unstructured and unprocessable for computers, manual processing of the reports is time-consuming and error-prone. Therefore, an automatic safety risk identification method is needed to address the challenge of processing a sizeable textual dataset.

This paper proposed a workflow to use the Text mining (TM) method, referred to as text data mining, to automatically identify critical safety risk factors hidden in accident reports. TM can discover valuable information and getting insights hidden in plain texts (Cheng et al. 2012). Different domains have their unique lexicon. For example, 'Shield' is known as a type of tunneling boring machine in underground construction; while, it generally refers to objects to protect a human from dangers. This paper also established a construction domain-specific lexicon, which plays a vital role in the TM workflow. Many terms are mentioned in the reports, to achieve more efficient and effective mining result, they need to be prioritized and reduced to a manageable size. This research proposed a method to evaluate term importance, which can reduce the impact of report length. Also, a threshold for identifying the high-frequency terms was defined to extract critical safety risk factors.

In summary, the core contributions of this research are:

- Devised a TM framework to extract critical risk factors in construction accident reports.

- Established a workflow for building a tailored domain lexicon.
- Proposed a novel method to evaluate the importance of terms in accident reports. The method integrates the Information entropy and term frequency (TF) and thus can reduce the impact of different report length.
- Proposed a quantified method to define the threshold of high and low frequency terms.

A case study of accident reports of metro construction projects in China is presented to illustrate the approach.

2. Literature review

2.1 Safety risk identification learning from past accidents

Accidents that occur, irrespective of the specific domain, have a strikingly similar trajectory (Dallat et al. 2019). Learning from past accidents has gained inspiration from research initiatives over the past few years. Simulation and optimization technics for safety risk assessment have advanced in the past 20 years (Alkaissy et al. 2020), such as Failure Mode and Effects Analysis (FMEA) (Ilbahar et al. 2018). However, safety risk identification in those models was limited to experience-based methods (e.g., literature review, questionnaires, etc.). Various accident causation theories and models were proposed based on the induction analysis of accidents, such as the Swiss Cheese model, the Man-Made Disaster Theory, the System-Theoretic Accident Model and Processes (STAMP), etc. (Yang and Haugen 2018). These theories have highlighted the primary mechanisms of how risk factors might cause an accident. However, the detailed safety risk factors were not clarified in the accident causation models.

Concerning safety risk factors, two traditional approaches have been used to identify them from past accidents. The first is a statistical analysis of accident data, using a pie chart, histogram, etc. For example, XU (2016) stated the time tendency and causes based on a statistical analysis of 167 metro construction accident reports; however, only one primary cause was considered per accident due to the sizeable manual work. Similarly, Zhou C et al. (2017) revealed temporal characters and dynamics of interevent time series of near-miss accidents by mapping time series into a complex network. This approach's predominant work is to transform the accident information into structured data by manual analysis or using structured data directly. Thus, it performs well at revealing the whole occurrence laws of workplace accidents (e.g., occurrence time, location, number of fatalities, accident types), but poor at extracting accident causes.

The second is a retrospective analysis of one or several accidents manually. For instance, Zhou Z and Irizarry (2016) conducted a detailed cause analysis of the foundation pit collapse accident in Hangzhou Metro. This approach provides a delicate analysis of causes but has sample limitations.

Through a preliminary literature review, it has been found that study on safety risk identification has little progress since the last decades. Dedicated research on identifying safety risk factors using the intensive resource is limited; this, in turn, conditions the risk evaluation and response. To address this, content analysis was proposed to seek out more productive results for safety risk identification from intensive accident cases (Esmaeili et al. 2015a, 2015b). Also, statistical analysis was utilized to reveal the accident causes and their characteristics based on a big database. For example, BİLİR and GÜRCANLI (2018) calculated the most frequently occurred accident types and construction jobs from 623 construction accidents, and provided the accident probabilities using activity-based accident rates and exposure values. KALE and Baradan (2020) developed a model to identify the factors that contribute to severity using a hybrid statistic technic, i.e., descriptive univariate frequency analysis, cross-tabulation, binary logistic regression. However, these methods still rely on expert's analysis to extract risk factors from texts. People use different expressions to describe similar factors. Factors may be ignored, misclassified, or merged by mistake. Therefore, the text mining method is proposed in this study to extract risk factors objectively from a large dataset of accident cases.

2.2 Risk identification using a text mining approach

TM refers to the process of extracting interesting, non-trivial information and knowledge from unstructured text documents that are not previously known and not easy to be revealed (Miner 2012). Eighty percent of construction data is stored in the text format (Ur-Rahman and Harding 2012). As for risk identification, studies have been conducted to extract useful information from text documents, such as contract risks from contract conditions (Siu et al. 2018), extracting socio-technical risks from licensee event reports of nuclear power plants (Pence et al. 2020). However, TM has rarely been used to identify safety risk factors from construction accident reports.

TM's primary step is to convert unstructured and semi-structured text to a structured format for further analysis (Jeehee and June-Seong 2017). Typical approaches include adaptive lexicon and natural language processing (NLP). The adaptive lexicon/dictionary method uses words predefined in a lexicon/dictionary to structuralize text. NLP transforms text into a semi-structured format with tags according to the sentence structure so that computers can understand. Machine-learning algorithms are generally used to improve the processing's effectiveness (e.g., artificial neural network)

(Ghosh and Gunning 2019). However, NLP methods usually require a large volume of domain-specific documents for training computers (Moon et al. 2019).

Figure 1 shows that structured data can be used in different ways to correspond with the aims of analysis. Researchers have used clustering and classification methods to categorize safety risks and link extraction methods to identify risk factors' inter-relationship. For example, Zhang F et al. (2019) proposed five baseline models: support vector machine (SVM), linear regression (LR), K-nearest neighbor (KNN), decision tree (DT), Naïve Bayes (NB), and an ensemble model to classify the causes of the accidents using the data from Occupational Safety and Health Administration (OSHA). Siu et al. (2018) proposed a classification approach to categorize the ordinary risks of the New Engineering Contract (NEC) projects to identify the critical risk factors. This paper only discusses the concept extraction methods, which aim to extract a list of risk factors - individual terms that already exist in the source documents - from the text.

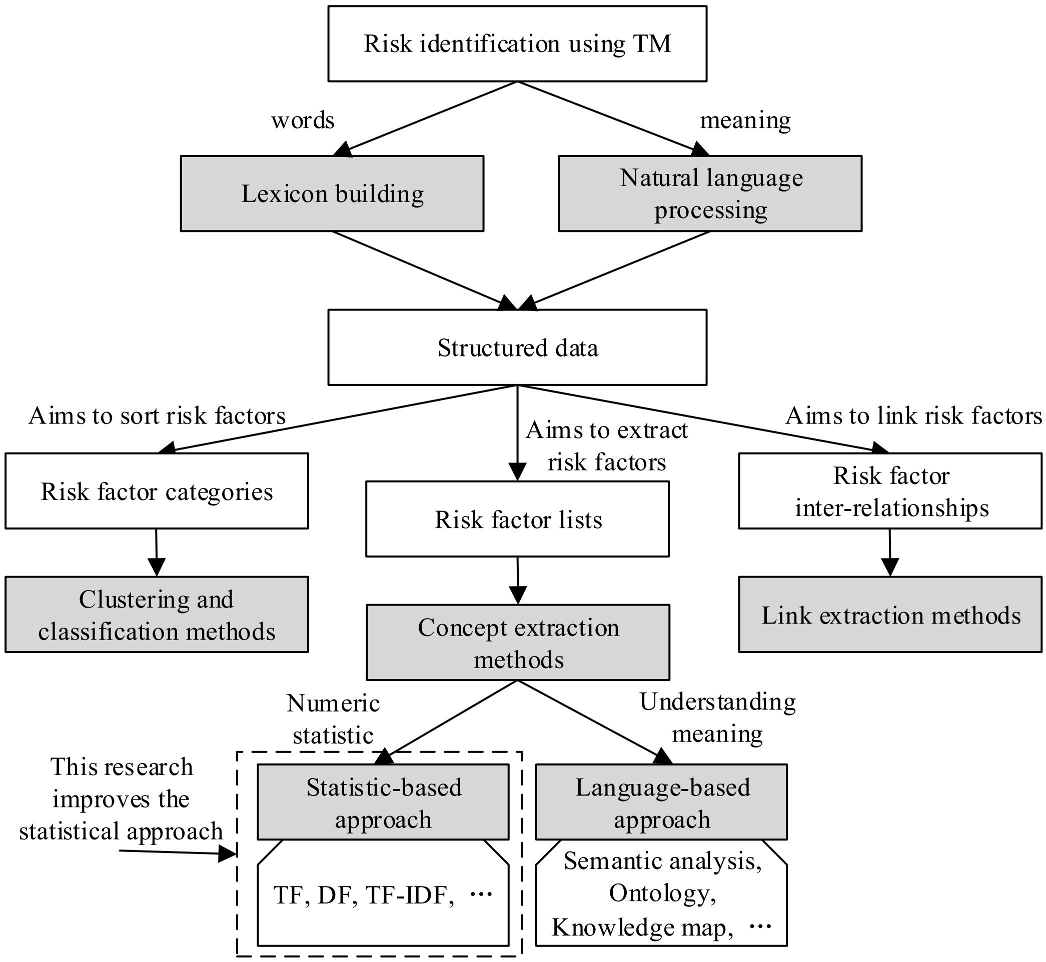


Figure 1. Risk identification using TM

Concept extraction methods (also called keyword extraction technology) mainly include the language-based and statistic-based approaches. The language-based approach uses semantic meanings and the rules of language structure to extract key terms. For example, Zhong et al. (2020) identified implied potential hazards comparing the annotations of construction site images with the specifications using semantic net and ontologies. This research uses a statistical approach to extract safety risk factors.

The statistical approach uses numeric statistics, such as TF, document frequency (DF), and term frequency-inverse document frequency (TF-IDF), to identify documents' features. For instance, Joon-Soo and Byung-Soo (2018) collected 10,798 internet news articles as a corpus; the most frequently occurred words (i.e., TF value) on fire-accidents were considered the most critical factors. Zhanglu et al. (2017) analyzed 41,791 hidden danger records of a coal mining enterprise, using a word cloud and TF to extract coal mine safety risks. Li et al. (2018) established a lexicon and used document frequency (i.e., DF value) and identified 15 high occurred safety risk factors and 3 participants from 156 accident reports. In Jeehee and June-Seong (2017), TF-IDF was utilized to prioritize the words from the prebid request for information (RFI) documents, and the mean value of TF-IDF was used to define the threshold of high-frequency terms. The detailed analysis will be provided in section 3.3.

Although some studies have made efforts to extract specific factors using high-frequency words from the text document, the method still needs to be improved according to different corpus and extracting aims. Also, the threshold for identifying critical factors, i.e., high-frequency terms, was commonly defined subjectively and needed to be improved.

3. Methodology

Figure 2 shows the framework of extracting safety risk factors from construction accident reports.

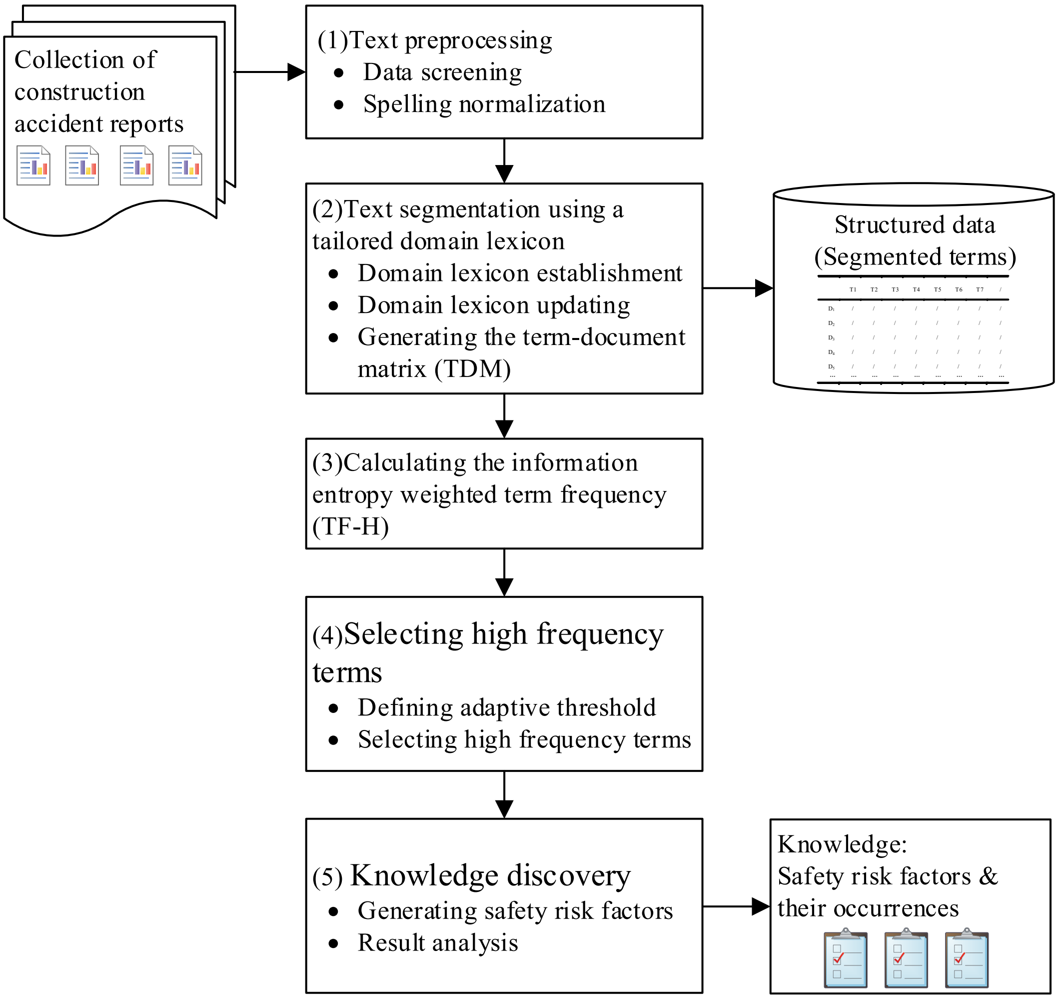


Figure 2 Framework for safety risk identification using TM approach

3.1 Text preprocessing

This step aims to clean and normalize the corpus, i.e., text-type construction accident reports. Two sub-steps, data screening and spelling normalization, are designed. Stemming, lemmatization, and case normalization are not needed for Chinese text preprocessing, making the text preprocessing different from the English text.

(1) *Data screening.* Remove the repeating and defect reports (e.g., incomplete reports).

(2) *Spelling normalization*. Unify misspellings, and spelling variations occurred in the corpus.

3.2 Text segmentation using a tailored domain lexicon

This step breaks the corpus into discrete and linguistically-meaningful terms (tokens) by locating the term boundaries, the points where one term ends and another begins (Miner 2012). Due to the diversities of human language, the descriptions of safety risk factors are of significant discrepancies. For example, 'rain' and 'storm' are probably used to describe similar weather conditions in the text; 'building firm' and 'construction company' both mean the 'contractor'. Therefore, to perform a better text segmentation, the dominating work is to construct a tailored domain lexicon.

Technically, the existing lexicon construction methods are mainly divided into corpus-based, knowledge-based methods, and their combination (Feng et al. 2018). Many domain words in the construction industry are specific phrases composed of common words, such as ‘construction management plan’ and ‘gantry crane.' It would be much easier to build the domain lexicon based on an existing common lexicon. Therefore, a combined method integrating corpus-based (use common lexicon to establish original domain lexicon) and knowledge-based (use experts' manual analysis to update domain lexicon) is designed in this study. Figure 3 shows the workflow of domain lexicon building, including lexicon establishment and lexicon updating.

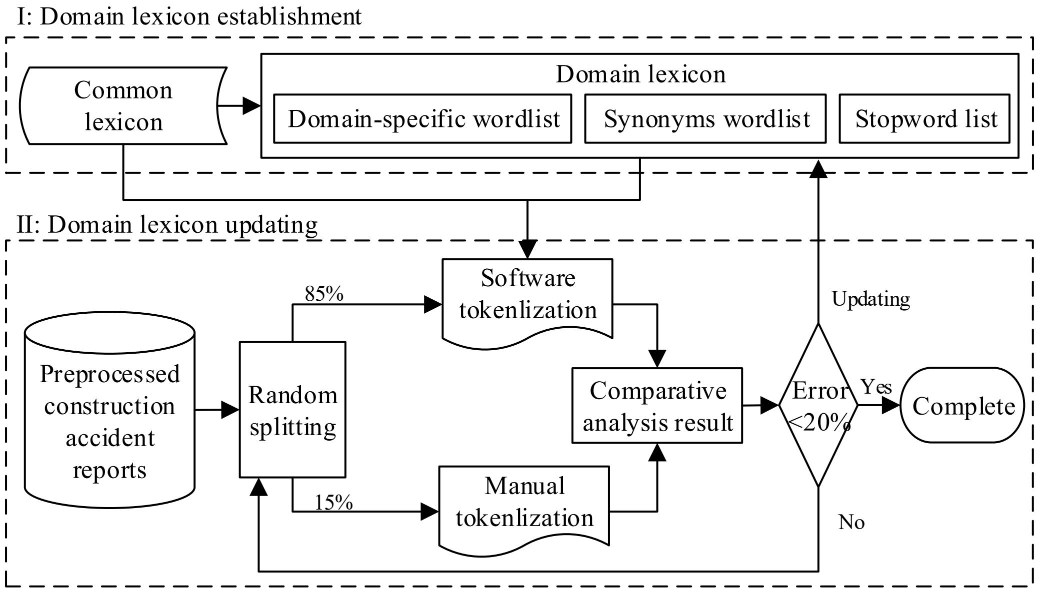


Figure 3. The workflow of domain lexicon building

3.2.1 Domain lexicon establishment

The following three wordlists are designed to be built in sequence.

(1) *Domain-specific wordlist*: Although most of the common words in the construction industry (e.g., *timber, tube, etc.*) has been encapsulated in the dictionaries of civil engineering, more domain-specific words are still in need, such as *shield, shaft, SMW (soil mixing wall), TBM (tunnel boring machine),* etc. Also, a set of common words may compose a phrase with specific meanings, such as *Diagonal bracing*, *horizontal bottom tube, foundation pit*, etc. Thus, the specific phrasal words need to be identified as one term instead of breaking them into meaningless single words.

(2) *Synonyms wordlist*: This wordlist aims to reduce the discreteness of language description and increase terms' frequency with the same meaning. For instance, *collapse*, *sloughing*, *collapsing*, and *fall* can all be replaced by *collapse*.

(3) *Stopword list*: Stopword refers to the word which appears in nearly every document while meaningless, such as *this* and *there*. Generally, they have only a grammatical function. These meaningless words need to be removed in order to highlight the effect of information extraction.

3.2.2 Domain lexicon updating

A computer processes 85% of the reports using a common lexicon while domain experts assess the rest for cross-checking. The two sets of results are compared. New words or phrases that are identified by experts but missed by the computer will be added to the lexicon. The computer gives preference to phrases. For example, if a new phrase 'construction management plan' is added to the domain lexicon, the whole phrase will be extracted when they occur together. The single word 'construction', 'management' and 'plan' will be extracted separately only when they occur alone. Therefore, the critical work of the domain lexicon building is to update new specific-matter words and phrases. The lexicon building process runs iteratively until the error rate is acceptable. The calculation of the error rate is shown in Eq. (1),

$$E = \frac{|\bar{A}|}{|A \cup B|} \tag{1}$$

where A refers to the set of terms tokenized by computer, B indicates the union set of terms identified by the domain experts, and $|A \cup B|$ means the number of elements in the union of A and B; $|\bar{A}|$ means the number of missing terms identified by experts but missed by computer. For instance, if $A = \{a, b, c, d\}$, $B = \{b, d, e, f\}$, then $\bar{A} = \{e, f\}$, and $E = 2/6 = 33\%$. The error rate is defined as $E = 20\%$, referring to Esmaeili et al. (2015a, 2015b) and Li et al. (2018).

3.2.3 Generating the term-document matrix

The segmented terms are vectorized into a sparse two-dimensional matrix, i.e., term-document matrix (TDM). TDM is a structured representation of the corpus, as shown in Eq. (2). Each column represents a term $t_i, i \in m$; each row represents a document $D_j, j \in n$; each cell's value represents how many times a term appears in a document called TF ($tf_{i,j}$). After that, the unstructured accident reports are converted to structured numerical data for further analysis.

$$TDM = \begin{bmatrix} tf_{1,1} & tf_{2,1} & tf_{3,1} & \cdots & tf_{m,1} \\ tf_{1,2} & tf_{2,2} & tf_{3,2} & \cdots & tf_{m,2} \\ tf_{3,1} & tf_{3,2} & tf_{3,3} & \cdots & tf_{m,3} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ tf_{m,1} & tf_{m,2} & tf_{m,3} & \cdots & tf_{m,n} \end{bmatrix} \tag{2}$$

3.3 Calculating the information entropy weighted term frequency (TF-H)

3.3.1 Traditional term-importance evaluation

The frequency of a term reflects its prominence to each report, i.e., the importance of a risk factor to each occurred accident. TF, DF, and $TF-IDF$ are the most widely used methods to evaluate term importance. Table 1 displays the comparison of the three methods.

Table 1. Traditional term-importance evaluation methods

Methods	Descriptions	Advantages	Limitations
$TF_{i,j}$	The frequency number of the term t_i appears in document D_j.	Reflects the total frequency count of a term.	Largely impacted by the length of reports.
D	The frequency number of documents that term t_i appears in the corpus.	Eliminates the impact of report length.	Lost the data of term frequency in one document.

$TF-IDF$	The comprehensive impacts of TF and inverse DF.	Consider the positive impact of TF and the negative impact of DF.	Not applicable to the occurrence features of safety risk factors.

Usually, the greater a term's TF value is, the greater the term contributes to this corpus. However, it cannot be said that safety risk factor A is more critical to accident I than accident II if the TF of term A in report I is higher than the TF of term A in report II. Some exceptions could be that report I is longer and more detailed; hence, A is mentioned more times. The impact of report length should be reduced or eliminated. Some studies used DF, meaning the number of documents containing the term, to represent the importance of risk factors (Li et al. 2018). However, the DF method leaves out the occurrence frequency that a term appears in the document. To address this, $TF-IDF$ was proposed to balance the impact of TF and DF. Inverse Document Frequency (IDF) means that the more frequently a term appears in all documents, such as ‘is’, the less it should weigh in a search (Zhang 2019). The calculation is shown in Eq. (3),

$$TF-IDF = tf_{i,j} \times idf_i \tag{3}$$

where $idf_i = \log\frac{|D|}{DF_i}$, $|D|$ is the total number of documents, DF_i is the document frequency containing the term t_i. $TF-IDF$ value is in direct proportion to TF and inversely proportional to DF. Therefore, $TF-IDF$ is often used to evaluate the critical feature of a document, i.e., a term can represent a document in the corpus in order to cluster the documents (Singh et al. 2019).

However, for the occurrence of safety risk factors, the more uniformly the term distributed in the accident report corpus, the more frequently the safety risk factor appears in different accidents, and more important should the factors be. None of the above methods has measured the document distribution of terms, which is very important for safety risk factors. Therefore, the priority of risk factors should be in direct proportion to the TF and the uniform distribution in the corpus.

3.3.2 Improved term-importance evaluation: TF-H

This research proposes $TF-H$ to evaluate the importance of a term to a document in the corpus. Information entropy (H), also known as Shannon entropy, is used to weigh the disorder's extent and its effectiveness in system information (Mohsen and Fereshteh 2017). Applied in risk evaluation techniques, the smaller the entropy value, the smaller the degree of dispersion of the index, and the greater the amount of information it carries, so the weight of this index in the system safety analysis is greater (Liu C et al. 2020). Therefore, the concept of information entropy reflects the occurring characteristic of risk factors. According to the information entropy formula, i.e., $H=-\sum p_i log p_i$, the $TF-H$ is defined as Eq. (4),

$$TF-H(t_i)=TF(t_i)\times H(t_i)=-tf_{i,j}\times\sum p_i log p_i \tag{4}$$

where p_i refers to the probability distribution of term t_i, $p_i=\frac{tf_{i,j}}{\sum_{j=1}^{n} tf_{i,j}}$; $H(t_i)$ characterizes the distribution of term t_i in the accident reports.

The proposed $TF-H$ method integrates the overall impacts of TF and the distribution of the term. With the information entropy of term distribution, the impact of report length can be largely reduced. Thus, compared to the other three traditional methods, $the\, TF-H$ method is more applicable for extracting essential terms representing safety risk factors.

3.4 Selecting high-frequency terms

To capture the critical safety risk factors, redundant data shall be filtered out. As the boundary between high and low-frequency terms, the adaptive threshold shall be well set. There are no given rules to define high-frequency words (Pang and Zhang 2019). One of the most popular methods is Donohue's formula $T = (-1 + \sqrt{1 + 8 \times I_1})/2$ (Donohue 1973), where T indicates the high-frequency word threshold; I_1 indicates the number of words that have only appeared once.

The TF, DF, or $TD - IDF$ was generally used to evaluate the term importance (YiShan et al. 2017). For example, Joon-Soo and Byung-Soo (2018) used cumulative TF to define the threshold, and terms less than 90% was removed. Pang and Zhang (2019) defined the keywords that appeared more than four times as high-frequency keywords. In this study, the accumulative $TF - H$ value is proffered to define the high-frequency term threshold based on the classical ABC grouping method. ABC method classifies the objects with accumulative values (Hasani and Mokhtari 2019).

Figure 4 shows the division of high-frequency terms based on the accumulative $TF - H$. The abscissa represents the segmented terms. The left ordinate represents the value of $TF - H$, while the right ordinate represents the accumulative $TF - H$ value. In order to achieve the accumulative $TF - H$ value, we need to convert the $TF - H$ value into the proportion form and then sort descending and obtain the accumulative sum. The terms in the interval of 0% to 90% are considered high-frequency terms (A-class), the rest as low-frequency terms.

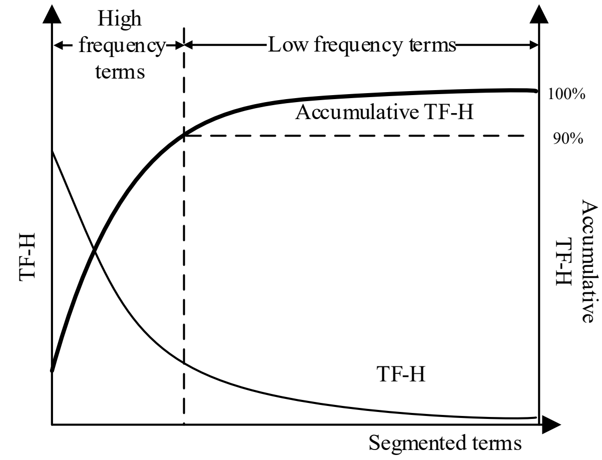


Figure 4. High-frequency term threshold based on accumulative TF-H value

(1) *High-frequency terms*: With the increase of the number of segmented terms, the $TF-H$ curve suddenly drops, and the accumulative $TF-H$ curve increases rapidly, indicating that the number of high-frequency terms is small, but the contribution to the overall corpus is significant, accounting for 90%;

(2) *Low-frequency terms*: With the increase of the number of segmented terms, the $TF-H$ curve slowly decreases, and the accumulative $TF-H$ curve increases slowly, indicating that the number of low-frequency terms is enormous, but the contribution to the overall corpus is small, only 10%.

3.5 Knowledge discovery

Contextualize the high-frequency terms in the accident reports and select the terms that indicate the safety risk factors (represented as S_i). Experts' knowledge is needed to match the high-frequency terms and safety risk factors to find valuable information.

4. Case study

Metro construction projects are subject to high safety risks due to the unpredictable geological conditions, complex construction methods, and surrounding construction conditions (Ding L and Zhou 2013). An incident can cause significant economic loss and massive casualties. For example, a tunnel collapse accident in the Foshan metro construction project in 2018 caused eleven deaths, one missing, and eight severely injured (MOHURD 2018; Zhou X-H et al. 2019). The process of risk identification is complex and large amounts of experts and financial resources are needed because metro construction is large-scale and specific-domain undertakings (Zhang S et al. 2019). A risk factor check list is helpful for the practitioners to identify .This study aims to find typical safety risk factors in metro construction projects based on hundreds of accident reports using the proposed framework shown in Figure 2.

4.1 Extracting safety risk factors using TF-H

Because metro construction has great social attention, there is much short news reporting the possible causes and injuries on websites. However, these reports are poor-quality, because they are released by non-professionals and contain little information. Therefore, we use the accident report that 1) is published by government authorities or written by professionals, and 2) has a plentiful description of the accident. Finally, two hundred twenty-one accident reports of metro construction projects were chosen as the corpus. They were acquired from: 1) websites of national and local administration of work safety, such as *Ministry of Housing and Urban Rural Development of the People's Republic of China* (MOHURD) and the *Ministry of Emergency Management of the People's Republic of China*, and 2) published papers and books for practitioners, and 3) and internal documents from metro construction enterprises. 68, 90 and 63 reports were collected from websites, publications and enterprises, accounting for 31%, 41%, and 28%. Table 2 shows the profile of data sources, and Figure 5 plots the geographic distribution of cities that accidents occurred. The accidents cover 27 cities (up to 80% of cities that run metro lines in China) from 1999 to 2017. The geographic distribution is concentrated in the east of China, because the eastern area is more developed. All the accident reports were stored as text files in a file folder for further processing.

Table 2. Profile of data sources

No.	City	Data Sources			Sum
		Websites	Publications	Enterprises	
1	Guangzhou	16	10	7	33
2	Shenzhen	13	7	10	30
3	Beijing	7	15	5	27
4	Shanghai	3	10	11	24
5	Wuhan	5	16	3	24
6	Nanjing	8	5	1	14
7	Qingdao	2	5	4	11
8	Xuzhou			9	9
9	Xi’an		1	5	6
10	Hangzhou	4	2		6
11	Dalian	1	3	1	5
12	Harbin	2	2	1	5
13	Fuzhou	1	3	1	5
14	Chengdu	1	1	1	3
15	Chongqing		3		3
16	Nanning	2		1	3
17	Ningbo	1		1	2
18	Kunming	1	1		2
19	Changchun			1	1
20	Shenyang		1		1
21	Tianjin	1			1
22	Xiamen		1		1

23	Zhengzhou		1		1
24	Wuxi		1		1
25	Lanzhou			1	1
26	Dongguan		1		1
27	Nanchang		1		1
	SUM	68(31%)	90(41%)	63(28%)	221

Figure 5. Geographic distribution of cities that accident occurred

Domain-specific wordlist was established based on the *Dictionary of civil engineering* downloaded from dictionaries in the *Google Input Method* and *Baidu Input Method*. Some words were defined with new meanings used in the specific domain, such as *shield, drainage*, and new phrases were added, such as *tunnel boring machine* and its abbreviation (TBM), *soil nailing support*, etc. Synonyms wordlist was established based on the *Dictionary of synonyms words (extended version)* developed by the Harbin Institute of Technology. For example, 'support system', 'support structure', 'bracing system', and 'bracing structure' were all represented by 'support system'. For stopwords, most of them can be found in the *Dictionary of Modern Chinese Function Words* downloaded from *Google Input Method* and *Baidu Input Method*. Besides, words that repeatedly appear in all reports but have no special meaning for analysis, such as *metro, accident, cause, process,* and *adopt*, were also added to the stopword list. One hundred eighty-eight reports (85% of the corpus) were processed by the computer, and the extracted tokens were composed of the set A in Eq. (1). Three experienced construction professionals conducted the manual tokenization to build the domain lexicon according to Figure 3. Table 3 shows the profile of the professionals. Thirty-three reports (15% of the corpus) were analyzed by them to extract the tokens, respectively. An in-depth discussion was conducted to reach an agreement on different tokens. Finally, the identified tokens composed the set B in Eq. (1). Then, the error rate E was calculated according to Eq. (1). The repeating process was carried out in four rounds, i.e., the terms in the domain lexicon were updated four times until the error was acceptable.

Table 3. Profile of the construction professionals

Code	Working years	Job title	Educational background	Department
A	20	Professor	Ph.D.	University
B	13	Project manager	Bachelor	Construction enterprise
C	25	Engineer	Master	Construction enterprise

Two thousand nine hundred ninety terms were obtained after text segmentation using the tailored domain lexicon, forming a TDM according to Eq. (2). The size of the full matrix is 221 by 2,990. Table 4 shows part of the TDM. For example, the segmented term T_1 appears once in the report document D_2, so $tf_{1,2}$ is 1; $tf_{9,6} = 21$ indicates that the term T_9 appears 21 times in the report document D_6.

Table 4. Term-document matrix

$tf_{i,j}$	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}	…	$T_{2,990}$
D_1	0	0	0	0	0	0	0	0	0	0	…	0
D_2	1	0	0	0	2	0	0	0	0	0	…	0
D_3	2	1	0	0	0	0	0	1	0	0	…	0
D_4	2	1	0	0	0	0	0	1	0	0	…	0
D_5	0	0	0	0	0	0	0	0	0	0	…	2
D_6	2	2	1	0	0	0	0	0	21	0	…	0
D_7	0	0	0	4	0	0	0	0	0	1	…	0
D_8	0	0	0	0	0	0	0	0	0	1	…	4
D_9	4	1	0	0	0	0	0	0	0	1	…	0
D_{10}	1	1	1	1	0	0	0	0	1	2	…	0
…	…	…	…	…	…	…	…	…	…	…	…	…
D_{221}	0	0	1	0	0	0	0	0	0	4	…	0

According to Eq. (4), the value of $TF-H$ was achieved. Subsequently, 253 high-frequency terms met the threshold (accumulative $TF-H \geq 90\%$) and were extracted. Table 5 shows the part of the high-frequency terms. The characteristics of construction workplace accidents are briefly highlighted. For example, 'foundation pit' and 'interval tunnels' indicate the section of metro construction; 'collapse' refers to the most frequent type of accidents (XU 2016); ‘construction enterprises’ implies the primary responsible party of workplace accidents. Finally, the high-frequency terms were traced back to the context in the reports; thirty-seven safety risk factors (S_i) were summarised, as shown in Table 6. The entire safety risk factors can be found in Table 7.

Table 5. High-frequency terms (part)

No.	Terms	TF-H	No.	Terms	TF-H	No.	Terms	TF-H
1	safety	989	11	personnel	305	21	underground hydrology	156
2	foundation pit	603	12	Inspection	295	22	facilities	152

No.	Terms	TF-H	No.	Terms	TF-H	No.	Terms	TF-H
3	collapse	408	13	process	274	23	monitor	141
4	support system	529	14	geological structure	257	24	construction technology	134
5	management	521	15	loose soil	236	25	operation	133
6	safety consciousness	470	16	construction personnel	204	26	Safety guarding	129
7	operation against rules	421	17	rain sewer pipe	196	27	supervision	126
8	work	373	18	safety management system	195	28	water and mud inrush	124
9	construction enterprises	336	19	construction project	178	29	collapse	123
10	interval tunnels	314	20	remediation	161	30	sedimentation	120

413 Table 6. Safety risk factors extracted from construction workplace accident reports

No.	High-frequency terms	TF-H	Context description in accident reports	Safety risk factors inducted
S1	Support system	529	As advanced support is not conducted, or the already conducted support has deficiencies, the support (enclosure) system experiences instability failure. For instance, the tunnel face is not timely sealed, and the support is not timely implemented after blasting.	Instability of the support system
S2	Management	521	Field safety supervision is ineffective, including ineffective field safety management, weak management, understaffed safety management, no administrators supervising construction operations, failing to correct potential safety hazards, etc.	Disordered field management
S3	Operation against rules	470	Contractors operate against rules, including violating construction schemes, rules, regulations, standard specifications, and other requirements. For instance, during the process of dismantling the supporting structure—bailey beam—of one Chongqing metro line in February 2016, indirect stress-bearing member bars of bailey beam are blindly cut, resulted in momentary instability and the collapse of bailey beam.	Construction operations against rules
…	…	…	…	…
S37	…	6	…	Improper selection of

4.2 Comparative study of term-importance evaluation

Table 7 compares the values of TF, DF, $TF-IDF$, and $TF-H$ of the safety risk factor S_i. Take S_{11}, S_{13}, S_{15} as an example for comparison. Although $TF(S_{11}) = TF(S_{15}) = 105$, the DF value of S_{11} is much higher, indicating that S_{11} caused more workplace accidents. Therefore S_{11} shall be preferentially selected as high-risk factors. However, $TF-IDF(S_{11})$ is much lower than $TF-IDF(S_{15})$, indicating that $TF-IDF$ does not apply to the extraction of safety risk factors from accident reports. Also, the DF value of S_{11} equals that of S_{13}, and $TF(S_{13}) > TF(S_{11})$. It seems that S_{13} should be more critical. However, the information entropy value shows that $H(S_{11}) = 1.45 > H(S_{13}) = 1.2$. This indicates that the distribution of S_{11} in accident reports is relatively uniform; namely, it has been mentioned multiple times in multiple accident reports, but S_{13} are mentioned several times in an accident report while less mentioned in other accident reports. Therefore, the importance of S_{11} is slightly higher than that of S_{13}. The above data comparison has favorably verified TF-H's superiority in measuring risk factors compared with traditional methods.

Table 7. Results of term-importance evaluation methods

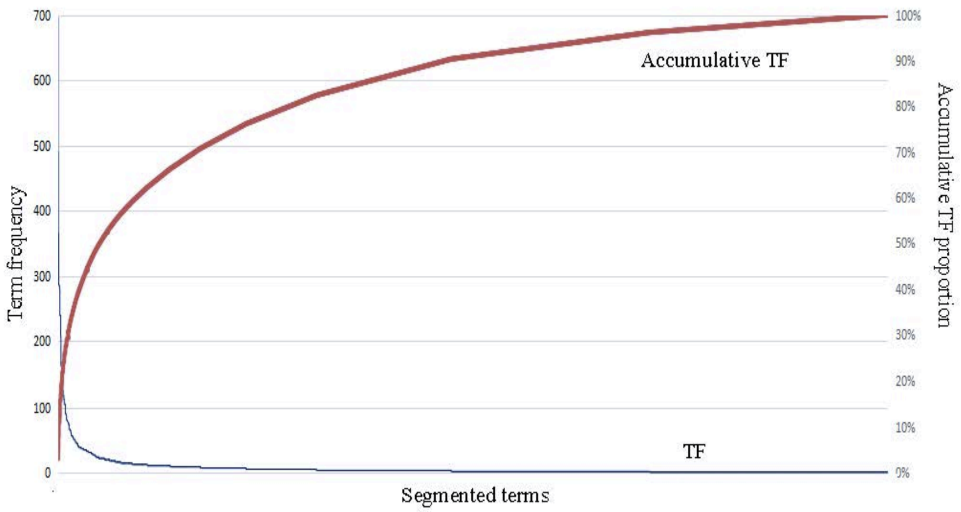
S_i	Safety risk factors	TF-H(S_i)	TF(S_i)	DF(S_i)	TF-IDF(S_i)	H(S_i)
S_1	Instability of the foundation pit support system	529.2	326	77	149.3	1.62
S_2	Disordered field management	521.1	319	79	142.5	1.63
S_3	Insufficient safety awareness	469.5	284	83	120.8	1.65
S_4	Construction operations against rules	420.6	282	81	122.9	1.49
S_5	Lack of safety inspection	294.8	184	74	87.4	1.6

S_i	Safety risk factors	TF-H(S_i)	TF(S_i)	DF(S_i)	TF-IDF(S_i)	H(S_i)
S_6	Complicated geological conditions	259.7	160	77	73.3	1.62
S_7	Insufficient exploration or protection of rain and sewage pipes	195.9	129	61	72.1	1.52
S_8	Ineffective safety management system	195.3	138	48	91.5	1.41
S_9	Insufficient remedial measures	160.6	111	52	69.8	1.45
S_{10}	Unclear underground hydrological conditions	156.4	103	61	57.6	1.52
S_{11}	Equipment and facility fault or inappropriate operation	152	105	52	66.0	1.45
S_{12}	Construction monitoring data lagging	141.1	120	55	72.5	1.18
S_{13}	Deficiency of construction technologies	133.7	111	52	69.8	1.2
S_{14}	Insufficient safety guarding	128.6	92	46	62.7	1.4
S_{15}	Dereliction of duty of the supervisor	126.4	105	29	92.6	1.2
S_{16}	Improper construction plan	117.3	85	44	59.6	1.38
S_{17}	Structural quality defect	110.6	85	37	66.0	1.3
S_{18}	Insufficient safety disclosure	108.2	92	28	82.5	1.18
S_{19}	Natural disaster	107.6	79	42	57.0	1.36
S_{20}	Insufficient exploration or protection of gas and power pipes	95	88	22	88.2	1.08
S_{21}	Lack of safety training	89.9	68	39	51.2	1.32
S_{22}	Lack of contingency plans and drills	87.2	64	42	46.2	1.36
S_{23}	Ineffective construction organization and coordination	84.6	64	39	48.2	1.32
S_{24}	Improper management of subcontractors	81	81	18	88.2	1
S_{25}	Construction not satisfying design requirements	76.6	61	33	50.4	1.26
S_{26}	Insufficient geological survey	61.5	50	33	41.3	1.23
S_{27}	Construction command against rules	45.3	42	22	42.1	1.08
S_{28}	Inappropriate crane hoisting or operation	44.2	41	22	41.1	1.08
S_{29}	Insufficient exploration or protection of surrounding buildings (structures)	30	30	18	32.7	1
S_{30}	Design defects	20.2	26	11	33.9	0.78

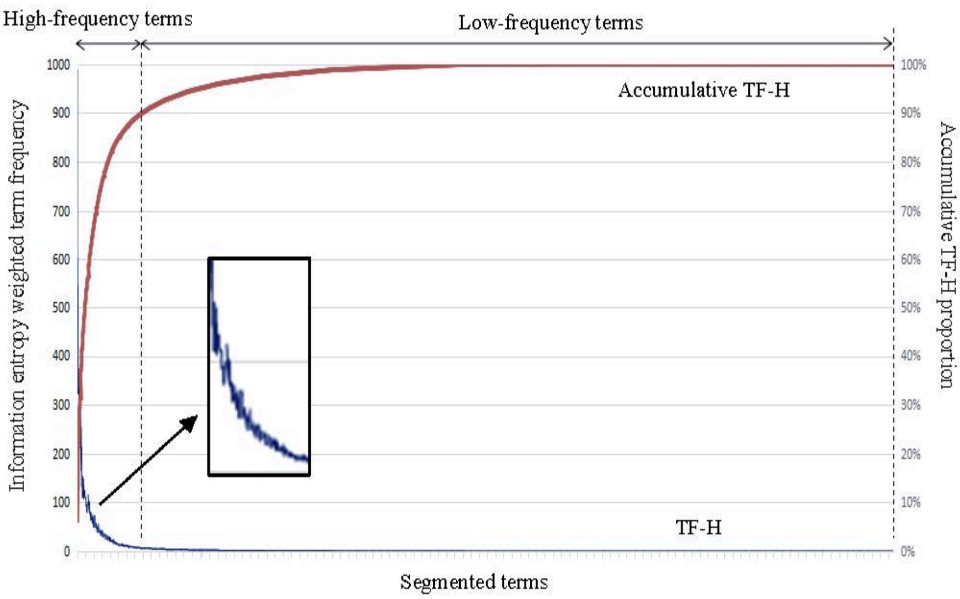
S_i	Safety risk factors	TF-H(S_i)	TF(S_i)	DF(S_i)	TF-IDF(S_i)	H(S_i)
S_{31}	Inappropriate goods and material placing	19.9	22	15	25.7	0.9
S_{32}	Pressure of construction period	10.6	13	7	19.5	0.82
S_{33}	Improper material selection	8.6	11	8	15.9	0.78
S_{34}	Defects of safety management organization	7.2	10	6	14.1	0.72
S_{35}	Form support system defects	7	10	6	15.7	0.7
S_{36}	Fatigue operation	6.8	9	6	14.1	0.75
S_{37}	Improper selection of mechanical equipment	6	8	6	12.5	0.75

4.3 Comparative study of threshold division

To test the effect of threshold division, two other methods were designed for comparative analysis, Donohue's formula and accumulative term frequency. Figure 5 compares the accumulated distribution of segmented terms from the perspective of TF and $TF-H$. Table 6 displays the results for selecting high-frequency terms using different methods.



(a) Term frequency (TF)



(b) Information entropy weighted term frequency (*TF-H*)

Figure 6. Accumulated distribution of segmented terms

Table 8. Comparison of high-frequency term selection methods

Methods	Threshold	Number of high-frequency terms
Donohue's formula	T=41	39
accumulative TF	≥90%	1401
accumulative TF-H	≥90%	253

Eight hundred sixty-one words only appeared once among all the tokens ($I_1 = 861$). Thus, the threshold $T = 41$, according to Donohue's formula described in Section 3.4. Donohue's formula depends on I_1. It can be seen from the TM distribution curve (Figure 6 (a)) that the number of terms that have appeared only once is large. Only 49 terms were selected, while 2,941 terms were filtered out. Therefore, this method may lead to massive missing items.

For the accumulative TF method, almost 50% of the terms were selected as high-frequency terms, resulting in the redundancy of words. This is because the accumulative TF curve (Figure 6 (a)) is smooth, the rise is slow, and there is no inflection point. Compared to the accumulative TF curve, the accumulative $TF-H$ curve (Figure 6 (b)) shows a rapid upward trend with a small number of segmented terms. There is a significant inflection point. Because the larger $the\,TF$ value of the term is distributed in the accident reports, the larger the information entropy will be. Therefore, the $TF-H$ value accelerates the rapid rise of the accumulation curve in the front part. Simultaneously, a large number of terms (including $TF=1$ and part $TF=2$ of the terms) in the long tail' have an information entropy of *0*, so that the accumulative $TF-H$ curve tends to be straight in the latter part. Therefore, compared to the accumulated TF value, the accumulative $TF-H$ value can better screen the high-frequency terms.

4.4 Result analysis of safety risk factors and their occurrences

4.4.1 Critical Safety risk factors of metro construction in China

High-frequency terms represent the critical safety risk factors of metro construction in China. According to Table 5, extracted safety risk factors mainly fall into the following five categories: surrounding environment, safety management, construction technology, construction personnel, materials, and equipment. Table 5 covers the main safety risk factors that Ding LY et al. (2012) and Xing et al. (2019) had mentioned.

Risk factors ‘instability of the foundation pit support system (S_1)’, ‘disordered field management (S_2)’, ‘insufficient safety awareness (S_3)', and 'construction operations against the rules (S_4)’ are the top four frequently occurred reasons leading to workplace accidents. Frequent inspection and monitoring of these factors are still necessary for the progressed metro projects to prevent similar accidents from happening.

‘Instability of the foundation pit support system (S_1)' is the most frequently occurred safety risk factors in metro construction projects. Most of the foundation pit support system is temporary. Thus, the construction company may take the chances to reduce the safety investment and shorten the construction time. Notably, a collapse accident may happen once S_1 is triggered, resulting in mass casualties. This confirms the conclusion in Liu et al. (2018) that the most significant risk factor in mechanical tunneling was improper soil reinforcement and drainage, and the main consequences included gushing water and collapse. However, in Liu et al. (2018), a large-scale questionnaire (514 responses) was conducted in five cities in China.

‘Disordered field management (S_2)’ demonstrates that ineffective safety management still widely exists in metro construction practice. According to accident causation theory, safety management is the root reason for accidents (Yang and Haugen 2018). Metro construction projects are always associated with volumes of intersection construction work and need high-standard and high-efficient safety management. ‘Insufficient safety awareness (S_3)’ is the third important factor identified in accident reports and is high referred to by academic paper (Fung et al. 2016; Maiti and Choi 2019). ‘Construction operations against the rules (S_4)’ refers to unsafe behavior on the construction site. Most construction workers in China come from migrant workers, and there is a shortage of personnel in terms of mobility, lack of professional training (Liu Q et al. 2020). Therefore, risks related to construction personnel are a big problem in metro construction projects.

4.4.2 Other valuable discoveries

The uncertainties of metro construction projects are largely related to the complex surrounding environment. Geological and hydrological conditions have been highly mentioned by scholars, such as in reference (Dong et al. 2018; Li et al. 2018). As in the accident report, 'Complicated geological conditions (S_6)' and 'Unclear underground hydrological conditions (S_{10})' are the sixth and tenth high-frequently referred reason causing an accident. This indicates that the two factors have attracted lots of concerns, both in theory and practice. However, other underground risks, such as 'Insufficient exploration or protection of rain and sewage pipes (S_7)', 'Natural disaster (S_{19})', 'Insufficient exploration or protection of gas and power pipes (S_{20})' and 'Insufficient exploration or protection of surrounding buildings (structures) (S_{29})', are less mentioned by academics. As a high-frequent reason, Factors S_7 and S_{19} (mainly refers to rain) usually cause soil erosion around the foundation pit, resulting in severe collapse accidents. In terms of S_{20} and S_{29}, they usually cause gas leakage, power blackout, or settlement of adjacent buildings, leading to adverse social impacts in the community.

Contingency planning and emergency management need to be enhanced. Notably, factors 'Insufficient remedial measures' (S_9) and 'Lack of contingency plans and drills' (S_{22}) are not the causes of accidents, but they are essential to prevent the expansion of accident losses. They are often mentioned in accident investigation reports, while they are generally ignored by most of the existing risk lists. Some studies have proposed contingency risks for bidding and contracts (Turskis et al. 2012; Jeehee and June-Seong 2017). However, there is still little research in the construction safety domain.

Preconstruction risks are not the main reasons causing an accident, yet they need to be noticed. Several studies have claimed the importance of design risks for safety construction (Hossain et al. 2018; Yuan et al. 2019). This study shows that most safety risk factors come from the construction phase, whereas three origins in the preconstruction phase, e.g. ‘Insufficient geological survey (S_{26})’ and ‘Design defects (S_{30})’. Both factors rank the low frequency.

Equipment and facility risks need increasing attention. Not many factors are related to construction materials and equipment. This reflects that construction materials and equipment are not the main reasons for metro construction accidents. However, the factor ‘Equipment and facility fault or inappropriate operation (S_{11})’ needs an increasing concern. With the widespread use of mechanical devices instead of man labor, the performance of mechanical equipment has become an increasing risk factor on the construction site.

The factor ‘Pressure of construction period (S_{32})’ and ‘Fatigue operation (S_{36})’ reveals the fact of a tight schedule of China's current metro construction situation. This also shows that safety may be sacrificed due to workload pressure.

Another discovery is that multiple causes led to construction accidents jointly. As shown in Table 7, the sum of the document frequency of the 37 safety risk factors is 1419, so the average number of risk factors causing the workplace accident is about 1419/221≈6.4. This confirms the accident causation theory that although only two or three factors cause workplace accidents directly, there is a wide range of risk factors hidden during the whole period of metro construction lifecycle, causing accidents indirectly.

5. Conclusion

Analyzing the workplace accident reports leads to learning from what went wrong in the past to prevent future accidents. An appropriate approach for text mining reduces the effort and increases the performance to discover valuable knowledge. This paper aims to provide an improved approach to extract safety risk factors effectively and efficiently from construction accident reports.

A text mining framework for safety risk factor extraction was proposed. A domain lexicon, including domain-specific wordlist, synonyms wordlist, and stopword list, was built to achieve a better text segmentation. An improved term-importance evaluation approach, $TF-H$, was provided to integrate the term frequency and the distribution of risk factors in accident reports. Accumulative $TF-H$, which was proposed to define the threshold to select high-frequency terms. This approach's improvement is that it introduces the distribution of a term in the corpus, and thus more applicable for the characteristic of safety risk factors. Then, a case study for safety risk factor extraction from metro construction accident reports was conducted. With the comparative analysis in the case study, the proposed approach was verified a better performance. The identified safety risk factors can comprehensively reflect the critical risks that metro construction projects encountered in China. Also, many interesting discoveries were found based on the implied information in the accident reports. The result will guide the practitioners to supplyment the safety risk factors of the project to be constructed, and avoid similar workplace accidents. The improved approach can also be used in other TM tasks to extract critical terms distributed in different lengths of documents.

Since the safety risk factors are extracted from accident reports, the information implied in the report determines the mining result. Many accident analysis studies have shown that risk factors can emerge outside the project, for example, local government, regulatory body, and social environment (Dallat et al. 2019; Lu et al. 2020). These latent outside risks are not included in accident reports but need to be noticed and assessed. Additionally, the mining result partly depends on experts' knowledge, including building domain lexicon and contextualizing the high-frequency terms. Manual intervention primarily lies in the inspection of computers' analysis to achieve a better result. Also, low-frequency terms were omitted as redundant data in this study because the computer extracted novel patterns by counting. Some low-frequency terms could be interesting for identifying new emerging risk factors. However, this will lead to much more redundant data and experts' knowledge to select.

Several possible future improvements can be considered. Extraction of valuable information from text documents differs given different corpus and tasks (Talib et al. 2016). More interesting results might be found if a broader corpus could be executed, such as journal papers, onsite documents, etc. Also, Different construction activities imply different safety risks, and different risks lead to different severities. More accident characteristics can be analyzed from the reports to reveal more mechanisms of workplace accidents, such as identifying the activity-based factors, the causal-and-effect relationship among factors, and the factor-and-severity relationship.

References

Alkaissy M, Arashpour M, Ashuri B, Bai Y, Hosseini R. 2020. Safety management in construction: 20 years of risk modeling. Safety science. 129:104805.

BİLİR S, GÜRCANLI GE. 2018. A Method For Determination of Accident Probability in Construction Industry. Teknik Dergi. 29(4):8537-8561.

Cheng C-W, Leu S-S, Cheng Y-M, Wu T-C, Lin C-C. 2012. Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry. Accident Analysis & Prevention. 48:214-222.

Dallat C, Salmon PM, Goode N. 2019. Risky systems versus risky people: To what extent do risk assessment methods consider the systems approach to accident causation? A review of the literature. Safety Science. 119:266-279.

Ding L, Zhou C. 2013. Development of web-based system for safety risk early warning in urban metro construction. Automation in Construction. 34:45-55.

Ding LY, Yu HL, Li H, Zhou C, Wu XG, Yu MH. 2012. Safety risk identification system for metro construction on the basis of construction drawings. Automation in Construction. 27:120-137.

Dong C, Wang F, Li H, Ding L, Luo H. 2018. Knowledge dynamics-integrated map as a blueprint for system development: applications to safety risk management in Wuhan metro project. Automation in Construction. 93:112-122.

Donohue JC. 1973. Understanding scientific literature: A bibliographic approach. Cambridge: The MIT Press.

Esmaeili B, Hallowell MR, Rajagopalan B. 2015a. Attribute-based safety risk assessment. I: analysis at the fundamental level. Journal of Construction Engineering and Management. 141(8):04015021.

Esmaeili B, Hallowell MR, Rajagopalan B. 2015b. Attribute-based safety risk assessment. II: predicting safety outcomes using generalized linear models. Journal of Construction Engineering and Management. 141(8):04015022.

Feng J, Gong C, Li X, Lau RYK. 2018. Automatic approach of sentiment lexicon generation for mobile shopping reviews. Wireless Communications and Mobile Computing. 2018:13.

Fung IWH, Tam VWY, Sing CP, Tang KKW, Ogunlana SO. 2016. Psychological climate in occupational safety and health: the safety awareness of construction workers in South China. International Journal of Construction Management. 16(4):315-325.

Ghosh S, Gunning D. 2019. Natural language processing fundamentals. Packt Publishing.

Gul M. 2018. A review of occupational health and safety risk assessment approaches based on multi-criteria decision-making methods and their fuzzy versions. Human and Ecological Risk Assessment: An International Journal. 24(7):1723-1760.

Gul M, Ak MF. 2018. A comparative outline for quantifying risk ratings in occupational health and safety risk assessment. Journal of Cleaner Production. 196:653-664.

Hasani A, Mokhtari H. 2019. An integrated relief network design model under uncertainty: A case of Iran. Safety Science. 111:22-36.

Hossain MA, Abbott ELS, Chua DKH, Nguyen TQ, Goh YM. 2018. Design-for-Safety knowledge library for BIM-integrated safety risk reviews. Automation in Construction. 94:290-302.

Ilbahar E, Karaşan A, Cebi S, Kahraman C. 2018. A novel approach to risk assessment for occupational health and safety using Pythagorean fuzzy AHP & fuzzy inference system. Safety Science. 103:124-136.

Jeehee L, June-Seong Y. 2017. Predicting project's uncertainty risk in the bidding process by integrating unstructured text data and structured numerical data using text mining. Applied Sciences. 7(11):1141.

Joon-Soo K, Byung-Soo K. 2018. Analysis of fire-accident factors using big-data analysis method for construction areas. KSCE Journal of Civil Engineering. 22(5):1535-1543.

KALE ÖA, Baradan S. 2020. Identifying Factors that Contribute to Severity of Construction Injuries using Logistic Regression Model*. Teknik Dergi. 31(2):9919-9940.

Karasan A, Ilbahar E, Cebi S, Kahraman C. 2018. A new risk assessment approach: Safety and Critical Effect Analysis (SCEA) and its extension with Pythagorean fuzzy sets. Safety science. 108(2018):173-187.

Li J, Wang J, Xu N, Hu Y, Cui C. 2018. Importance degree research of safety risk management processes of urban rail transit based on text mining method. Information. 9:26.

Liu C, Yang S, Cui Y, Yang Y. 2020. An improved risk assessment method based on a comprehensive weighting algorithm in railway signaling safety analysis. Safety Science. 128:104768.

Liu Q, Xu N, Jiang H, Wang S. 2020. Psychological Driving Mechanism of Safety Citizenship Behaviors of Construction Workers: Application of the Theory of Planned Behavior and Norm Activation Model. Journal of Construction Engineering and Management. 146(4):04020027.

Liu W, Zhao T, Zhou W, Tang J. 2018. Safety risk factors of metro tunnel construction in China: An integrated study with EFA and SEM. Safety Science. 105:98-113.

Lu L, Li W, Mead J, Xu J. 2020. Managing major accident risk from a temporal and spatial perspective: A historical exploration of workplace accident risk in China. Safety Science. 121:71-82.

Maiti S, Choi J-h. 2019. An evidence-based approach to health and safety management in megaprojects. International Journal of Construction Management.1-13.

Miner G. 2012. Practical text mining and statistical analysis for non-structured text data applications. 1st edition ed. Academic Press.

Mohsen O, Fereshteh N. 2017. An extended VIKOR method based on entropy measure for the failure modes risk assessment – A case study of the geothermal power plant (GPP). Safety Science. 92:160-172.

MOHURD. 2018. Accident letters. China: Ministry of Housing and Urban- Rural Development of the People's Republic of China [accessed]. http://sgxxxt.mohurd.gov.cn/Public/AccidentList.aspx.

Moon S, Lee G, Chi S, Oh H. 2019. Automatic Review of Construction Specifications Using Natural Language Processing. ASCE International Conference on Computing in Civil Engineering 2019; 2019; Atlanta, Georgia.

Nonaka I. 2008. The knowledge-creating company. Harvard Business Review Press.

Pang R, Zhang X. 2019. Achieving environmental sustainability in manufacture: A 28-year bibliometric cartography of green manufacturing research. Journal of Cleaner Production. 233:84-99.

Pence J, Farshadmanesh P, Kim J, Blake C, Mohaghegh Z. 2020. Data-theoretic approach for socio-technical risk analysis: Text mining licensee event reports of U.S. nuclear power plants. Safety Science. 124:104574.

PMI. 2017. A Guide to the Project Management Body of Knowledge (PMBOK guide). PA 19073 USA: Project Management Institute; 6th ed edition (30 Sept. 2017).

Qazi A, Quigley J, Dickson A, Kirytopoulos K. 2016. Project Complexity and Risk Management (ProCRiM): Towards modelling project complexity driven risk paths in construction projects. International Journal of Project Management. 34(7):1183-1198.

Rivas T, Paz M, Martín JE, Matías JM, García JF, Taboada J. 2011. Explaining and predicting workplace accidents using data-mining techniques. Reliability Engineering and System Safety. 96(7):739-747.

Singh K, Maiti J, Dhalmahapatra K. 2019. Chain of events model for safety management: Data analytics approach. Safety Science. 118:568-582.

Siu M, Leung W, Chan W. 2018. A data-driven approach to identify-quantify-analyse construction risk for Hong Kong NEC projects. Journal of Civil Engineering and Management. 24(8):592-606.

Soliman E. 2018. Risk identification for building maintenance projects. International Journal of Construction Project Management. 10(1):37-54.

Talib R, Hanif MK, Ayesha S, Fatima F. 2016. Text mining: techniques, applications and issues. International Journal of Advanced Computer Science and Applications. 7(11):414-418.

Tembo-Silungwe C, Khatleli N. 2018. Identification of enablers and constraints of risk allocation using structuration theory in the construction industry. Journal of Construction Engineering and Management. 144(5).

Turskis Z, Gajzler M, Dziadosz A. 2012. Reliability, Risk Management, and Contingency of Construction Processes and Projects. 18(2):290-298.

Ur-Rahman N, Harding JA. 2012. Textual data mining for industrial knowledge management and text classification: A business oriented approach. Expert Systems with Applications. 39(5):4729-4739.

Xing X, Zhong B, Luo H, Li H, Wu H. 2019. Ontology for safety risk identification in metro construction. Computers In Industry. 109:14-30.

XU N. 2016. Occurrence Tendency and Cause Analysis of Safety Accidents in Rail Transit Projects. Journal of Huaqiao University (natural edition). 37(5):6.

Yang X, Haugen S. 2018. Implications from major accident causation theories to activity-related risk analysis. Safety Science. 101:121-134.

YiShan L, YuLin W, MingXin L. 2017. An Empirical Analysis for the Applicability of the Methods of Definition of High-Frequency Words in Word Frequency Analysis. Digital Library Forum. 9:42-49.

Yuan J, Li X, Xiahou X, Tymvios N, Zhou Z, Li Q. 2019. Accident prevention through design (PtD): Integration of building information modeling and PtD knowledge base. Automation in Construction. 102:86-104.

Zhang F, Fleyeh H, Wang X, Lu M. 2019. Construction site accident analysis using text mining and natural language processing techniques. Automation in Construction. 99:238-248.

IEEE, editor. The Research on General Case-Based Reasoning Method Based on TF-IDF. 2019 2nd International Conference on Safety Produce Informatization (IICSPI); 28-30 Nov. 2019 2019; Chongqing, China. IEEE.

Zhang S, Shang C, Wang C, Song R, Wang X. 2019. Real-Time Safety Risk Identification Model during Metro Construction Adjacent to Buildings. Journal of Construction Engineering and Management. 145(6):04019034.

Zhanglu T, Xiao C, Qingzheng S, Xiaoci C. 2017. Analysis for the potential hazardous risks of the coal mines based on the so-called text mining. Journal of Safety and Environment. 17(4):1262-1266.

Zhong B, Li H, Luo H, Zhou J, Fang W, Xing X. 2020. Ontology-based semantic modeling of knowledge in construction: classification and identification of hazards implied in images. Journal of Construction Engineering and Management. 146(4):04020013.

Zhou C, Ding L, Skibniewski MJ, Luo H, Jiang S. 2017. Characterizing time series of near-miss accidents in metro construction via complex network theory. Safety Science. 98:145-158.

Zhou X-H, Shen S-L, Xu Y-S, Zhou A-N. 2019. Analysis of Production Safety in the Construction Industry of China in 2018. Sustainability. 11(17):4537.

Zhou Z, Irizarry J. 2016. Integrated Framework of Modified Accident Energy Release Model and Network Theory to Explore the Full Complexity of the Hangzhou Subway Construction Collapse. Journal of Management in Engineering. 32(5):05016013.