ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20231213001200
Original scientific paper

Data-driven Analysis of Risk Factors for Coal Mine Accidents

Yang YANG, Zhilei WU, Fenfen SHI*

Abstract: Coal mine safety production management is always essential for economic and social security and stability, analyzing and exploring the causes of coal mine accidents and the intrinsic correlation between various factors can effectively realize the control and containment of coal mine accidents. Existing research on cause analysis and risk assessment in the coal mine field mostly depends on a limited number of experts rather than data. This study uses text mining methods to extract historical coal mine accident information and integrates DEMATEL, ISM, and MICMAC methods to analyze coal mine safety accidents in a data-driven manner. This paper analyzes the fundamental, direct, and key factors of coal mine accident risk, reveals the causal relationship of accidents, and proposes relevant prevention and control measures based on the characteristics of each factor. The research results show that the data-driven risk analysis model can eliminate traditional methods' dependence on experts and provide theoretical references for preventing coal mine accidents.

Keywords: coal mine accidents; DEMATEL-ISM; risk analysis; text mining

1 INTRODUCTION

Coal is one of the most widely used and important energy sources in the world. In 2023, global coal demand exceeded a record high of 8.5 billion tons for the first time. Safe coal mining is an important foundation for supporting global energy security and economic stability. China is the world's largest producer and consumer of coal [1]. According to statistics, the number of deaths caused by coal mine accidents in China has exceeded 60000 since 2000 [2]. Coal mine accidents have become the biggest challenge to coal mine safety production. Effectively utilizing modern information technology to strengthen the collection, mining, and correlation analysis of safety production big data, enhance the ability to identify and evaluate safety risks, is of great significance in curbing the occurrence of major accidents and ensuring production safety needs. In recent years, many scholars at home and abroad have researched the risk assessment of coal mine accident causation. Wang et al characterized and statistically analyzed the more significant and above coal mine accidents from 2011 to 2020 and analyzed the causes of the typical cases of accidents that occurred in coal mine production systems from the four aspects of man, machine, environment and management with the help of the theoretical knowledge of safety entropy [3]; Li et al. carried out the risk assessment of gas explosion in coal mines by using the fuzzy Bayesian network [4]; Zhang et al [5] analyzed the causative factors of gas explosion and established the Bayesian network model of gas explosion from the point of view of the fault tree. Li [6] and Zhang [7] used accident trees and hierarchical analysis to establish a water penetration accident evaluation model to analyze the causes of water penetration accidents and comprehensive evaluation. Meanwhile, association algorithms and data mining techniques have been applied to coal mine safety management in recent years. In order to effectively assess the risk of gas explosion in coal mines, Cheng Lian et al. proposed a gas explosion risk assessment method based on explanatory structural modeling (ISM) and Bayesian network [8]. Lei et al. [9] used a correlation algorithm to data mine the causal factors of coal mine gas and obtained the causal chain of gas accidents. It can be seen that some progress has been made in existing research, but most studies have limited data or rely on expert scoring for risk assessment, which has strong subjectivity and lacks in-depth research on the interaction and coupling relationship between influencing factors and the degree of impact of each factor on accidents. The development of technologies such as text mining and information extraction has provided new ideas for risk analysis in the coal mining field. Over the years, in the process of coal mine safety management in China, a large number of accident case investigation reports, hidden danger investigation ledgers, and other materials that can extract risk information have been accumulated. These pieces of information are important basis for analyzing and preventing accidents. It is crucial to use natural language processing technology to mine key accident information from these unstructured data, identify the causal chain of accidents for effective analysis, and prevent accidents from occurring. Based on this feature of text mining, in recent years, it has been gradually used to explore the causes of security accidents in complex systems (Raviv et al., and Singh et al.)[10, 11]. For example, Gao and Wu [12] developed a verb-based text mining method to extract the causes and results of 945 automobile traffic accidents from the web version of the traffic accident reports, which helped to understand the real cause of the traffic accident. Nayak et al. [13] analyzed the main causes of urban traffic accidents based on the captured accident data report, indicating the importance of traffic facility optimization and road route planning. Jia et al. [14] proposed an improved decision- making test and evaluation laboratory (KG-CN-DEMATEL) based on knowledge graphs and complex networks, using knowledge graphs with Gaussian embedding (KG2E) to vectorize risk-related textual information. Bai et al. [15] proposed a novel risk assessment model integrating Knowledge Graph, DEMATEL and BN to analyze natural gas pipeline accidents in a data-driven manner to reduce reliance on experts. Liu et al. [16] used natural language processing (NLP) and text mining technology to mine pipeline accident data to understand the influencing factors and causes behind the incident. Text mining can help researchers deeply understand and discover factors and associations that affect safety production, thereby improving the efficiency and accuracy of accident

prevention. However，the application in the field of coal mine safety is still in the exploratory stage. In summary, there is currently a lack of effective methods for in-depth mining and utilization of massive unstructured text data on coal mine safety accidents. This article attempts to introduce text mining technology, scientifically integrate risk management theory with modern information technology, and analyze a large number of text reports on coal mine accident cases to study the efficient and accurate identification of coal mine safety risk factors. It provides a data-driven method for analyzing coal mine safety risk factors. In this study, the information extraction method was first used to identify the causes and chains of 838 typical coal mine accidents. Based on this, 22 coal mine safety risk causal factors were extracted and coded at level 3. Then, DEMATEL is applied to calculate each factor's influence degree, influenced degree, cause degree, and center degree to quantify the complex correlation in the causal network. Finally, ISM's method is used to construct a multilevel recursive order structure model for coal mine safety risk factors and combined with the MICMAC method to classify the types of risk factors and analyze the influence degree of risk factors at each level. The results of this study can better elucidate the evolutionary characteristics of coal mine accidents and effectively support risk prevention and safety management of coal mine accidents.

2 METHODS

In this paper, optimization research is carried out based on DEMATEL-ISM method and integration of MICMAC method, forming the DEMATEL-ISM-MICMAC coal mine safety risk analysis method, which analyzes the risk factors of accidents from the data-driven perspective, proposes safety risk prevention and control measures, and provides coal mine safety management with reliable decision-making support, the specific flow chart is shown in Fig. 1.

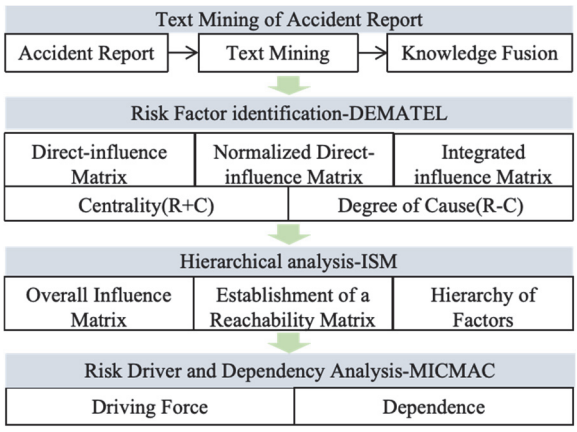

Figure 1 DEMATEL-ISM-MICMAC method based on text mining

2.1 Text Mining

According to global IDC (Internet Data Center) statistics, the proportion of unstructured data within existing enterprises is as high as 80%, and it will grow exponentially at a rate of 60% annually [17]. Exploring and obtaining the value hidden behind unstructured data such as text is of great significance in curbing the occurrence of major accidents and ensuring safety production requirements. Text mining is the process of extracting potentially important patterns or knowledge that users are interested in from unstructured text information. The mining of textual information is mainly based on mathematical statistics and computational linguistics, discovering the patterns of certain characters and the connections between characters, semantics, and grammar. Text mining can help researchers gain a deeper understanding and discover the factors and associations that affect safety production, thereby improving the efficiency and accuracy of accident prevention. However, its application in the field of coal mine safety is still in the exploratory stage. This study utilizes text mining techniques to effectively analyze the text of coal mine accident cases, extracting complete, effective, understandable, and valuable knowledge.

2.2 Dematel

Decision Making Experimentation and Evaluation of Experimental Methods (DEMATEL) is a system analysis method proposed using graph theory and matrix tools for solving complex and difficult system problems in the real world. DEMATEL is a structured and practical technique for causal networks to quantify correlations and dependencies between factors [18]. By assessing the significance and correlation of individual factors, DEMATEL can distinguish causal relationships, determine rankings and weights, and support decision making [19, 20]. The specific steps of DEMATEL are shown below.

Step 1: The coal mine safety risk elements summarized in the knowledge graph are labeled F_1, F_2, ..., F_n in order.

Step 2: A data-driven determination of the direct influence matrix is derived based on the constructed knowledge graph. By counting the frequency of occurrence of each correlation, the direct impact matrix $O=\left[O_{ij}\right]_{m\cdot n}$ in Tab. 1 can be obtained.

$$B=\begin{bmatrix} 0 & O_{12} & \cdots & O_{1n} \\ O_{21} & 0 & \cdots & O_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ O_{n1} & O_{n2} & \cdots & 0 \end{bmatrix} \tag{1}$$

Step 3: Analyze the indirect influence relationships of the factors in the system and normalize he direct influence matrix O to obtain the normalized direct influence matrix $Z=\left[Z_{ij}\right]_{m\cdot n}$.

$$Z=\frac{O}{\max\sum_{i=1}^{n}O_{ij}} \tag{2}$$

In the formula, $0 \le z_{ij} \le 1$, and $\max\sum_{j=1}^{n} z_{ij}=1$.

Step 4: Calculate the integrated impact matrix $T\left(T=\left[t_{ij}\right]_{n\cdot n}\right)$ according to the formula. In the formula, the factor t_{ij} in the matrix T represents the combined influence level of factor i on factor j, including the direct and indirect influence levels. The calculation formula is shown below.

$$T=\frac{z}{1-z} \tag{1}$$

In the formula, z is the unit matrix.

Step 5: Calculate the degree of influence a_i, the degree of being influenced b_i, the degree of center M_i, the degree of cause N_i for each element in the system. The specific calculation formula is as follows.

$$a_i=\sum_{i=1}^{n}t_{ij} \qquad i=1,\ 2,\ ...,\ n \tag{2}$$

$$b_i=\sum_{i=1}^{n}t_{ij} \qquad i=1,\ 2,\ ...,\ n \tag{3}$$

$$M_i=\sum_{i=1}^{n}t_{ij}+\sum_{i=1}^{n}t_{ji} \qquad i=1,\ 2,\ ...,\ n \tag{4}$$

$$N_i=\sum_{i=1}^{n}t_{ij}+\sum_{i=1}^{n}t_{ji} \qquad i=1,\ 2,\ \cdots,\ n \tag{5}$$

According to the calculation results, the centrality degree M_i is ranked, and the key risk factors can be determined. Risk factor attributes are determined by the positivity and negativity of the cause degree N_i. Finally, the result factors and cause factors are derived.

2.3 ISM

The Interpretative Structural Modelling (ISM) method is a widely used system science method, a model proposed by Professor Warfield, an American economist, in 1973 when he explored complex economic structures. The ISM method lists the influencing factors of the system to be studied and draws a directed graph based on the relationships between the influencing elements. Through Boolean logic operations, the ambiguous hierarchy of factors and the complex system composition are transformed into a clear ISM model [21]. DEMATEL and ISM have both commonality and high complementarity. The commonality lies in the fact that both methods use matrix theory, directed graph theory, and the analysis is based on a large number of mathematical theories. First, the multi-layer hierarchical explanatory structure model and DEMATEL analysis of the indicators of the degree of influence, the degree of influence, which can effectively circumvent the ISM can only analyze whether there is influence between two indicators, but cannot indicate the degree of influence between the two indicators of the shortcomings. Secondly, ISM reachable matrix can be calculated by DEMATEL's comprehensive impact matrix, which contains much more information than a single ISM reachable matrix, and the information is more detailed and comprehensive. It can form a more reliable multi-layer recursive explanatory structure model, which can decompose the complex system into a simple system. In this paper, based on the DEMATEL method, the ISM model is established and the calculation steps are simplified as follows.

Step 1: Take the integrated matrix T derived from the analysis of the DEMATEL method as the basis, and add it with the unit matrix to derive the overall system impact matrix $H = T + I$, where I is the unit matrix.

Step 2: Compute the reachable matrix $K=\left(k=\left[k_{ij}\right]_{n\cdot n}\right)$ according to Eq. (8).

$$K_{ij}=\begin{cases}0, & h_{ij}<\lambda\\ 1, & h_{ij}\geq\lambda\end{cases} \tag{8}$$

In the formula, λ is a set threshold, and the size is compared with the factor h_{ij} in the overall influence matrix H.

Step 3: Hierarchical division of reachable matrix K. Divide the factors in the reachable matrix K into reachable set $R(Fi)$ and prior set $Q(Fi)$. The reachable set $R(Fi)$ represents the set of elements in a reachable matrix or directed graph that are reachable by Fi. The precedence set $Q(Fi)$ denotes the set consisting of all other elements in the reachable matrix or directed graph in which Fi is reachable.

Step 4: Calculate and verify according to Eq. (9), if the formula is valid, the rows and columns belonging to it are delimited in the matrix K. The following steps are performed.

$$R\left(Fi\right)\cap Q\left(Fi\right)=R\left(Fi\right)\left(i=1,\ 2,\ ...,\ n\right) \tag{9}$$

Step 5: Repeat steps 3 and 4 to delimit all the factors in the system. Finally, the hierarchical relationship of factors is established in the order in which the factors are delineated.

2.4 MICMAC

In 1993, Duperrin et al. introduced the method of cross-matrix multiplication (MICMAC) to explore the diffusivity of interrelationships among the factors within the system based on the ISM model to classify the factors into different types. By calculating the dependencies and driving forces of the factors within the system, the key elements of the system were derived from the comprehensive analysis. The main calculation formula is as follows.

$$E_i=\sum_{i=1}^{n}k_{ij}\left(i=1,\ 2,\ ...,\ n\right) \tag{6}$$

$$F_i=\sum_{i=1}^{n}k_{ij}\left(i=1,\ 2,\ ...,\ n\right) \tag{7}$$

Based on the formula, drive and dependence were

calculated separately and the factors were categorized as autonomous (I), dependent (II), linked (III) and independent (IV).

3 RESULTS

3.1 Text Mining of Coal Mine Accidents

This study selects 2036 coal mine accident investigation reports that occurred in China from 2005 to 2021 as the data source, and retains 838 typical cases for text mining after screening and preprocessing. In order to obtain the causal chain that led to coal mine accidents, this article focuses on extracting disaster causing factors such as personnel, equipment, management, environment, and their causal relationships from the coal accident report text.

Due to the lack of a unified reporting format for coal mine accident reports, there is a high degree of uncertainty in the expression of their text, making NLP more difficult and greatly reducing the feasibility of text mining methods. In response to this issue, this article adopts a strategy based on rule matching, combined with PLM (Pre-training Language Model) and LLM (Large Language Model) for information extraction. In response to the textual content with distinct regular characteristics in accident reports, we designed regular expression templates to extract key elements through template matching. For professional vocabulary in the field of coal mine safety, we manually annotated data and pre-trained the UIE unified extraction model for extraction [22].

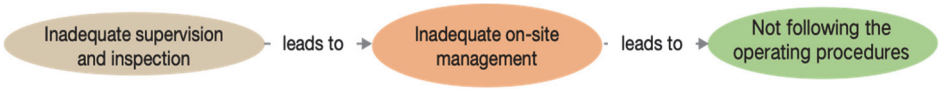

Figure 3 Causal relationship of coal mine disaster factors

Table 1 Spindle coding and selective coding

Category	Main Category		Initial Category (Part of the factors)
Management Factors	*F*1	Inadequate safety management system	Incomplete gas management system; Chaotic ventilation management system; No operation procedure
	*F*2	Inadequate safety supervision / inspection	Inadequate on-site inspections; Weak regulatory forces; Weak government supervision
	*F*3	Inadequate safety training and education	Inadequate safety training and education; Lack of targeted safety training; Outdated content
	*F*4	Inadequate implementation of prevention measures/safety responsibilities	Avoiding law enforcement inspections by regulatory authorities; No advance detection conducted
	*F*5	On-site management confusion	Unreasonable layout of the working face; Lack of strict on-site management; Unreasonable labor organization
	*F*6	Inadequate investigation and management of hidden dangers	Insufficient inspection and investigation of hidden dangers; Inadequate identification of safety risks; Insufficient efforts in flood investigation and control; Inadequate daily maintenance by electricians
	*F*7	Insufficient safety investment	Insufficient investment in safety costs; The side ditch is not covered with a cover plate; The safety device of the scraper conveyor is not properly equipped
	*F*8	Inadequate technical management	Adopting prohibited coal mining methods; The technical solution is unreasonable; Unreasonable gas extraction
	*F*9	Lack of safety responsible personnel	No safety guardian; Insufficient allocation of safety management personnel; No temporary security manager
	*F*10	Design defect	No mining design; There are defects in the support design of the operating procedures; Failure to predict the impact of fully mechanized mining on the mining face
	*F*11	Illegal production	Illegal production; Cross-border mining; Catch-up production
	*F*12	Poor emergency rescue measures	Inadequate organization of emergency rescue work; Insufficient rescue equipment and technical means
Personnel Factors	*F*13	Weak safety awareness	Insufficient awareness of safety risks; Lack of self-protection awareness
	*F*14	Illegal operations	Command in violation of regulations; Adventure homework; Unauthorized pinching of the air duct
	*F*15	Improper operation	Miner operation error; Unavoided mining trucks in operation
	*F*16	Unlicensed employment	Work without a certificate; Non full-time locomotive drivers drive electric locomotives at will; Driving a forklift without a license
	*F*17	Lack of skills, experience or knowledge	Low skills and techniques; Improper operation; Missing program steps; Insufficient emergency response capability; Insufficient safety knowledge
	*F*18	Bad physical and mental state	Lack of concentration; Insufficient work energy
Equipment Factors	*F*19	Unqualified equipment or poor reliability	Unstable ventilation system; The power supply equipment is unqualified; Unqualified pumping equipment; Unqualified monitoring equipment; Defective protective device
	*F*20	Equipment failure	Monitoring and monitoring equipment failure; Fatigue damage to the drum shaft; Inadequate equipment operation and maintenance
	*F*21	Missing equipment	No gas monitoring system installed; Insufficient number of sensors; No safety monitoring system; No portable gas detector; Missing Road signs and signs
Environmental Factors	*F*22	Complex / abnormal environment	Unclear hydrogeological conditions; The geological conditions of gas are complex; Working face fault

For common-sense knowledge, we utilized the Baichuan large language model to leverage its existing prior knowledge and reasoning capabilities for entity extraction, thereby reducing the cost of manual annotation. This paper employs a method of semantic similarity calculation to effectively aggregate various risk factors with different expressions in accident report texts. The concept for risk factor information extraction is as shown in Fig. 2, and one of the causal relationships of coal mine disaster factors is as shown in Fig. 3.

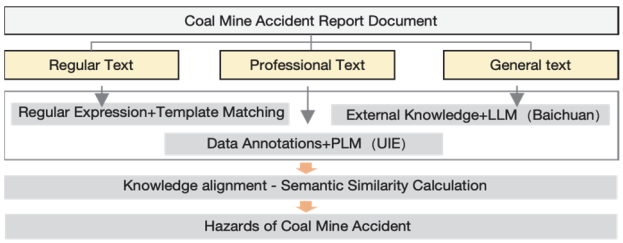

Figure 2 Design concept for risk factor information extraction

3.2 Coal Mine Accident Risk Factor Identification

Rooting theory, first proposed by Glaser and Strauss in the mid-20th century, is a qualitative research method that empirically generalizes the original data and then rises to the system [23], which mainly includes the processes of open coding, spindle coding, selective coding and theory saturation test. Combined with modern accident causation theory and system theory, the initial category, main category and core category are extracted by 3-level coding for the data source of the knowledge graph of coal mine accidents, such as the causative factors and cause analysis, to complete the identification of the causative factors of gas explosion. Through summarization, de-emphasis and generalization, the initial categories of the knowledge graph entities are coded in the main axis, resulting in 22 main categories; after systematic analysis and selective coding of the central categories, the causal factors of coal mine accidents are categorized into four core categories, namely, human factors, equipment factors, environmental factors and management factors. The results of principal axis coding and selective coding are shown in Tab. 1.

4 DISCUSSION

4.1 Direct Impact Matrix

The typical DEMATEL relies on linguistic opinions from experts during the determination of the direct-influence matrix. Drawing on Yiping Bai's method, this paper derives a data-driven determination of the direct impact matrix of the constructed knowledge graph by counting the frequency of occurrence of each correlation. This is done by statistically analyzing the number of directed arcs for each risk factor of coal mine safety accidents to derive the direct influence matrix ($O=\left[O_{ij}\right]_{23\cdot23}$) of coal mine risk factors. In the matrix O, the element O_{ij} indicates the degree of direct influence of the factor F_i on F_j. For example, $F_{13} = 6$ indicates that the directed arcs from F_1 to F_3 appear six times. If $i = j$ then $S_{ij} = 0$.

4.2 Integrated Impact Matrix

According to Eq. (2), the normalized direct impact matrix is obtained. Eq. (3) is then used to obtain the integrated impact matrix.

4.3 Center Degree and Cause Degree

According to Eq. (4) to Eq. (7), the influence degree a_i, the influenced degree b_i, the center degree M_i, the cause degree N_i of each risk factor are calculated, and the center degree is sorted, and the results are shown in Tab. 2.

Table 2. Causality indicators for all factors

Fi	Degree	Influenced degree	Center degree	Cause degree	Centrality ranking	Factor attribute
*F*1	0.834	0.085	0.918	0.749	10	cause
*F*2	1.341	0.240	1.581	1.102	3	cause
*F*3	1.586	0.312	1.898	1.274	1	cause
*F*4	0.932	0.506	1.438	0.425	4	cause
*F*5	0.512	0.559	1.072	–0.047	6	result
*F*6	0.637	0.657	1.294	–0.020	5	result
*F*7	0.347	0.386	0.733	–0.039	12	result
*F*8	0.553	0.485	1.038	0.068	7	cause
*F*9	0.365	0.281	0.645	0.084	16	cause
*F*10	0.202	0.234	0.435	–0.032	20	result
*F*11	0.076	0.400	0.475	–0.324	18	result
*F*12	0.037	0.471	0.508	–0.434	17	result
*F*13	1.197	0.455	1.652	0.743	2	cause
*F*14	0.043	0.943	0.986	–0.900	8	result
*F*15	0.019	0.678	0.696	–0.659	14	result
*F*16	0.248	0.445	0.693	–0.197	15	result
*F*17	0.570	0.376	0.946	0.194	9	cause
*F*18	0.179	0.200	0.379	–0.021	22	result
*F*19	0.121	0.581	0.702	–0.460	13	result
*F*20	0.017	0.716	0.733	–0.699	11	result
*F*21	0.002	0.456	0.459	–0.454	19	result
*F*22	0.035	0.389	0.425	–0.354	21	result

According to the causality index of risk factors in Tab. 2, the scatter plot of cause-results of risk factors of coal mine safety accidents is drawn, as shown in Supplementary Table S3 online. In Fig. 4, the x-axis represents the center degree and the y-axis represents the cause degree, where the cause factors are above the coordinate axis, representing that these factors directly affect the occurrence of coal mine safety accidents. The result factors are below the coordinate axis, and these factors are influenced by the cause factors, which indirectly affect the occurrence of safety accidents. The larger the value of the center degree, the greater the importance of the factors.

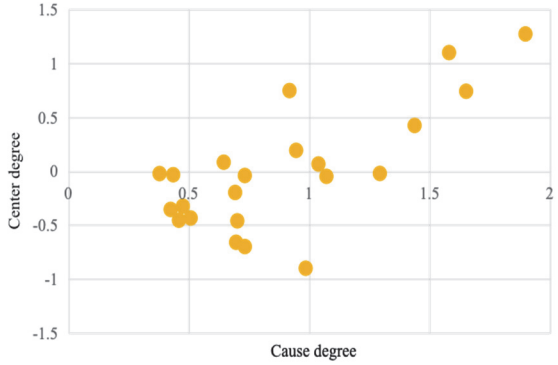

Figure 4 Causal scatter plot of risk factors

As can be seen from Tab. 2 and Fig. 4, the centrality rankings of coal mine safety accident risk factors are, in descending order, Inadequate safety training and education (*F*3), Weak safety awareness (*F*13), Inadequate safety supervision / inspection (*F*2), Illegal production (*F*11), Inadequate implementation of prevention measures/safety responsibilities (*F*4), Inadequate investigation and management of hidden dangers (*F*6), On-site management confusion (*F*5), Inadequate technical management(*F*8), Illegal operations (*F*14), Lack of skills, experience or knowledge (*F*17), Inadequate safety management system (*F*1), Equipment failure (*F*20), Insufficient safety investment (*F*7), Unqualified equipment or poor reliability (*F*19), Improper operation (*F*15), Unlicensed employment (*F*16), Lack of safety responsible personnel (*F*9), Poor emergency rescue measures (*F*12), Illegal production (*F*11), Missing equipment (*F*21), Design defect (*F*10), Complex / abnormal environment (*F*22), Bad physical and mental state (*F*18). The top five factors in terms of centrality are located in the upper right-hand side of the causal scatter plot.

4.4 Reachability Matrix Calculation

The method of converting from DEMATEL model to ISM model is that the reachability matrix can be calculated based on the integrated influence matrix T and the threshold value λ. Specifically, the overall impact matrix of the system is first calculated by Eq. (3). In recent years, some researchers began to use the average value of the comprehensive impact matrix as the threshold of the reachable matrix. For example, Lin Yan et al. used this method as the threshold of the overall relationship matrix when studying the influencing factors of the cost of steel structure prefabricated buildings, and achieved satisfactory model results [24]. Some scholars also make some deformation on the basis of the average method, such as adding the standard deviation of the sample to obtain the summation result and then use it as the threshold value. This method usually requires a small deviation and belongs to the optimization adjustment of the average method under specific circumstances [25-27]. The average value method can be used to influence the establishment of the relationship model when there are many factors, and at the same time, it can avoid a lot of early data collection work and improve the operability of the method. Therefore, the average value of the comprehensive influence matrix is chosen as the threshold of the reachable matrix. In the paper, the threshold λ is set with reference to the research of Lin Yan and other scholars, and the threshold in this paper is set to $\lambda = 0.066$. Finally, the reachable matrix K is calculated by using Eq. (8).

4.5 Division of Hierarchical Structure

Based on the reachability matrix, the matrix region is divided by the third step of the ISM method, which performs interval decomposition and inter-level decomposition of the matrix. Interval decomposition is to divide the elements into individual subsystems, and interlevel decomposition is to divide the elements within the same system into different hierarchies. The specific divisions are shown in Tab. 3 to Tab. 7.

Table 3 Division of elements in the first level

Fi	*R*(*Fi*)	*Q*(*Fi*)	*C*(*Fi*)
*F*1	1,4,5,8	1	1
*F*2	2,4,5,6,7,8,11,14,16,21	2	2
*F*3	3,4,5,6,8,11,13,14,15,16,17,20	3	3
*F*4	4,5,6,7,11,14	1,2,3,4	4
*F*5	5,6	1,2,3,4,5,13	5
*F*6	6,14,20,21,22	2,3,4,5,6,13	6
*F*7	7,9,21	2,4,7,13	7
*F*8	8,10,14	1,2,3,8	8
*F*9	9,12	7,9	9
*F*10	10,19,20	8,10	10
*F*11	11	2,3,4,11	11
*F*12	12	9,12,17	12
*F*13	5,6,7,13,14,16,17,18	3,13	13
*F*14	14	2,3,4,6,8,13,14,16,17	14
*F*15	15	3,15,16,17,18	15
*F*16	14,15,16	2,3,13,16	16
*F*17	12,14,15,17	3,13,17	17
*F*18	15,18	13,18	18
*F*19	19,20	10,19	19
*F*20	20	3,6,10,19,20	20
*F*21	21	2,6,7,21	21
*F*22	22	6,22	22

Table 4 Division of elements of the second level

Fi	*R*(*Fi*)	*Q*(*Fi*)	*C*(*Fi*)
*F*1	1,4,5,8,	1	1
*F*2	2,4,5,6,7,8,16	2	2
*F*3	3,4,5,6,8,13,16,17	3	3
*F*4	4,5,6,7	1,2,3,4	4
*F*5	5,6	1,2,3,4,5,13	5
*F*6	6	2,3,4,5,6,13	6
*F*7	7,9	2,4,7,13	7
*F*8	8,10	1,2,3,8	8
*F*9	9	7,9	9
*F*10	10,19	8,10	10
*F*13	5,6,7,13,16,17,18	3,13	13
*F*16	16	2,3,13,16	16
*F*17	17	3,13,17	17
*F*18	18	13,18	18
*F*19	19	10,19	19

Table 5 Division of elements in the third level

Fi	*R*(*Fi*)	*Q*(*Fi*)	*C*(*Fi*)
*F*1	1,4,5,8	1	1
*F*2	2,4,5,7,8	2	2
*F*3	3,4,5,8,13	3	3
*F*4	4,5,7	1,2,3,4	3,4
*F*5	5	1,2,3,4,5,13	5
*F*7	7	2,4,7,13	7
*F*8	8,10	1,2,3,8	8
*F*10	10	8,10	10
*F*13	5,7,13	3,13	13

Table 6 Division of elements in the fourth level

Fi	*R*(*Fi*)	*Q*(*Fi*)	*C*(*Fi*)
*F*1	1,4,8,	1	1
*F*2	2,4,8	2	2
*F*3	3,4,8,13	3	3
*F*4	4	1,2,3,4	3,4
*F*8	8	1,2,3,8	8
*F*13	13	3,13	13

Table 7 Division of elements in the fifth level

Fi	*R*(*Fi*)	*Q*(*Fi*)	*C*(*Fi*)
*F*1	1	1	1
*F*2	2	2	2
*F*3	3	3	3

As shown in Tab. 3 to Tab. 7, the first layer elements of coal mine safety accident risk factors are divided into $L1 = \{11, 12, 14, 15, 20, 21, 22\}$. Then the rows and columns where $F11$, $F12$, $F14$, $F15$, $F20$, $F21$, $F22$ are located are deleted, and the reachable and prior sets are divided to obtain the next layer of elements of this system model. Similarly, the second layer elements are obtained as $L2 = \{6, 9, 16, 17, 18, 19\}$. The third layer elements are $L3 = \{5, 7, 10\}$. The fourth layer elements are $L4 = \{4, 8, 13\}$. and the fifth layer elements are $L5 = \{1, 2, 3\}$ Finally, the system model is divided into five layers, $L = \{L1, L2, L3, L4, L5\}$. As a result, a multi-layer multilayer recursive order structure model of coal mine safety risk factors can be constructed as shown in Fig. 5. The first layer ($L1$) of factors belongs to the surface layer of direct factors, which is the direct factor causing coal mine safety accidents. They mainly include Illegal production ($F11$), Illegal operations ($F14$), Improper operation ($F15$), Poor emergency rescue measures ($F12$), Equipment failure ($F20$), Missing equipment ($F21$), Complex / abnormal environment ($F22$). The second ($L2$), the third ($L3$) and the fourth($L4$) layers of factors belong to the middle layer of indirect factors, which usually have an impact on the surface layer of direct factors, and are also affected by the underlying fundamental factors. The main factors include Weak safety awareness ($F13$), Inadequate implementation of prevention measures/safety responsibilities ($F4$), Inadequate technical management ($F8$), Insufficient safety investment ($F7$), On-site management confusion ($F5$), Design defect ($F10$), Inadequate investigation and management of hidden dangers ($F6$), Lack of safety responsible personnel ($F9$), Lack of skills, experience or knowledge ($F17$), Bad physical and mental state ($F18$), Unlicensed employment ($F16$), Unqualified equipment or poor reliability ($F19$).

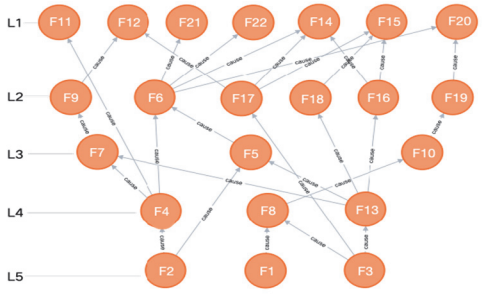

Figure 5 Multi-layer stepwise structural model

The fifth layer ($L5$) of factors belongs to the bottom of the fundamental factors, which will have a long-term impact on the upper layers of the system, and is not negligible and needs to be considered as a key factor, including Inadequate safety management system ($F1$), Inadequate safety supervision / inspection ($F2$), Inadequate safety training and education ($F3$).

4.6 MICMAC Analysis

Based on Eq. (10) and Eq. (11), the driving force and dependence were calculated respectively, and the factors were categorized into autonomy (I), dependence (II), linkage (III) and independence (IV). The results are shown in Tab. 8.

Table 8 Table of drive-dependence values

Fi	Risk factors	Driving	Dependency
F1	Inadequate safety management system	4	1
F2	Inadequate safety supervision / inspection	10	1
F3	Inadequate safety training and education	12	1
F4	Inadequate implementation of prevention measures/safety responsibilities	6	4
F5	On-site management confusion	2	6
F6	Inadequate investigation and management of hidden dangers	5	6
F7	Insufficient safety investment	3	4
F8	Inadequate technical management	3	4
F9	Lack of safety responsible personnel	2	2
F10	Design defect	3	2
F11	Illegal production	1	4
F12	Poor emergency rescue measures	1	3
F13	Weak safety awareness	8	2
F14	Illegal operations	1	9
F15	Improper operation	1	5
F16	Unlicensed employment	3	4
F17	Lack of skills, experience or knowledge	4	3
F18	Bad physical and mental state	2	2
F19	Unqualified equipment or poor reliability	2	1
F20	Equipment failure	1	5
F21	Missing equipment	1	4
F22	Complex / abnormal environment	1	2

Based on the results in Tab. 8, a driver-dependency classification diagram of coal mine safety risk factors was drawn, as shown in Fig. 6.

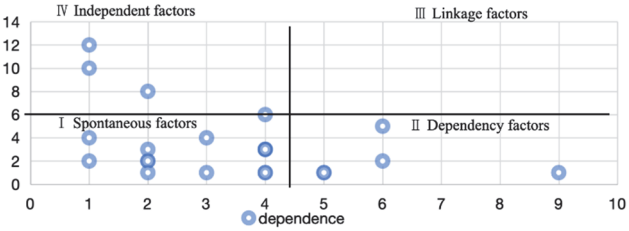

Figure 6 Driving-dependency classification

Factors within the first quadrant (I) are autonomous factors. Factors in this quadrant have low dependence and drive, which are mostly located in the middle layer of the multi-layer hierarchical structural model and play the role of connecting the top and the bottom. Factors in this quadrant mainly include Insufficient safety investment ($F7$), Inadequate technical management ($F8$), Lack of safety responsible personnel ($F9$), Design defect ($F10$), Illegal production ($F11$), Poor emergency rescue measures ($F12$), Weak safety awareness ($F13$), Unlicensed employment ($F16$), Lack of skills, experience or knowledge ($F17$), Bad physical and mental state ($F18$), Unqualified equipment or poor reliability ($F19$), Missing equipment ($F21$), Complex / abnormal environment ($F22$). Factors in the second quadrant (II) are dependent factors, which have a higher dependence but lower driving force, and are generally located in the middle and upper levels of the multilayered hierarchical structural model, with a lower degree of influence on other factors. Factors within On-site management confusion ($F5$), Inadequate investigation and management of hidden dangers ($F6$), Illegal operations

(*F*14). Improper operation (*F*15), Equipment failure (*F*20). Within the third quadrant (III) are linkage type factors. This type of factor has both high driving force and dependency, and is more unstable. The reason for the occurrence of this type of factor may be due to improper selection of factors or mingling of underlying and surface factors. The absence of linkage type factors in this article indicates that the factors selected in the article are stable. Factors in quadrant (IV) are independent type factors. Factors in this quadrant are more strongly driven but less dependent, and other factors have less influence on them. They are generally at the bottom of a multilayered hierarchical structural model and are fundamental factors in the model that continue to deeply influence the system. Factors within this quadrant are Inadequate safety supervision / inspection (*F*2), Inadequate safety training and education (*F*3), Weak safety awareness (*F*13), Inadequate implementation of prevention measures/safety responsibilities (*F*4).

4.7 Preventive Control Measures

In summary, the article constructs the data-driven model method to analyze the coal mine safety risk factors, which is simplified by the knowledge graph network, summarized as the direct complex matrix of the risk factors of coal mine safety accidents and formed the normalized influence matrix by using the normalization of the DEMATEL method, and based on which the comprehensive influence matrix is obtained. Meanwhile, the coal mine safety is analyzed by using the ISM method. Safety risk factors are divided into hierarchical structures, supplemented by the MICMAC method for driving and dependence analysis, and the following conclusions are drawn. First of all, the bottom factors in the risk factors of coal mine safety accidents are ineffective safety supervision/ inspection, inadequate safety training and teaching, and unsound safety management system, and they are all hidden factors with relatively high driving force and low dependence. This shows that the most fundamental factors of coal mine safety accidents mainly lie in the constructing of a safety management system, safety supervision, safety education, etc. Therefore, a governmental safety supervision and management organization with clear responsibilities and comprehensive coverage should be constructed, the coordination and communication of various departments and agencies should be strengthened, the relevant rules and regulations of supervision and management should be formulated, and the law should be strictly enforced. Enterprises should also improve the coal mine safety management system, refine the content of coal mine safety management, and effectively implement safety management. In addition, government regulators, coal mines, and other related units should strengthen safety education and improve the safety knowledge and awareness of managers and operators, which is the fundamental logic to avoid accidents. Secondly, the most direct risk factors for coal mine safety accidents include illegal operation, improper operation, equipment failure, lack of equipment, lack of emergency rescue measures and environmental heterogeneity/ abnormalities. These factors are mainly unsafe human behavior and unsafe state of things. Coal mine safety management focuses on production site management as the priority to avoid problems; site management is divided into the management of personnel and equipment management, in which personnel management should focus on the staffing situation personnel state of mind, according to the rules of operation and other aspects of the management of equipment should focus on the quality of equipment purchases, equipment operation and maintenance of the norms and systems equipped with a complete. In addition, coal mines should also pay attention to the construction of the spatial environment of coal mine production to avoid the hazards brought by environmental factors. Finally, the indirect risk factors in the middle layer of coal mine safety have strong driving forces and low dependence, and most of them belong to spontaneous factors, indicating that the factors in the middle layer are more stable. They mainly need low safety awareness, lack of knowledge, chaotic management, inadequate prevention and control measures, design defects, etc. It can be seen that the intermediate-level factors are mainly related to the implementation of the management of the enterprise; such factors are more volatile and at the same time a large number of factors have a wide range of influences, so it is the core factors affecting the risk of accidents. For the enterprise, it is necessary to strengthen the internal management of the enterprise, for example, to improve the safety awareness, to enhance the knowledge of mine safety, to implement the primary responsibility of the enterprise.

5 CONCLUSIONS

This paper used text mining technology to analyze the investigation report of coal mine accidents in China, and extracted the risk factors affecting safe production in a data-driven manner. The integrated DEMATEL-ISM methodology provided a data-driven characterization of coal mine accident risk factors. The text mining results elicited multi-layered dependencies and causal chains from historical cases. Clustering addressed imbalanced distributions. The ISM was utilized to delineate the systematic hierarchical structure of coal mine accident risk factors, and the MICMAC method was introduced to analyze the drivers and dependencies of risk factors. On this basis, the fundamental factors, direct factors and critical factors of the risk factors of coal mine accidents were analyzed, and the causal relationship before and after the accidents was revealed. The quantified interrelationships and hierarchical analysis inform targeted improvements in safety management, supervision, training, and monitoring to mitigate predominant risk drivers. The study demonstrates a broadly adaptable framework for uncovering systemic risk insights from unstructured data. Finally, we proposed the preventive and control measures based on the characteristics of each factor and suggested the directions for accident control and prevention. Obviously, in order to more effectively use a large number of accident reports to mine potential risk factors that should be paid attention to in a complex system, in addition to further optimizing the use of text mining, it is also necessary to establish a common accident report compilation standard.

Acknowledgements

This work was supported by the National Social Science Fund of China (Grant No. 22BGL110).

6 REFERENCES

[1] Pan, L. H., Zhao, P. P., Gong, D. L., Yan, H. M., & Zhang Y. J. (2022). Research on entity recognition method for naming coal mine accident cases. *Computer Technology and Development, 32*(02), 154-160.

[2] Xu, P. F. (2022). Study on characteristics and occurrence rules of coal mine accidents in china from 2000 to 2021. *Coal Engineering, 54*(07), 129-133.

[3] Wang, Y. G., Cui C. Y., Zhang, F. Y., Yao, K., & Kang J. B. (2022). Statistical analysis and research on major and above coal mine accidents in china from 2011 to 2020. *Journal of Safety and Environment*, 1-9.

[4] Li, M., Wang, H. T., Wang, D., M., Shao, Z. L., & He, S. (2020). Risk assessment of gas explosion in coal mines based on fuzzy AHP and bayesian network. *Process Safety and Environmental Protection, 135*, 207-218. https://doi.org/10.1016/j.psep.2020.01.003

[5] Zhang, J. J., Xu, K. L., Wang, B. B., & Wang Y. T. (2016). Extraordinarily serious gas explosion accidents in coal mines: analysis of causes and research on management mode. *China Safety Science Journal, 26*(2), 73-78. https://doi.org/10.1088/1475-7516/2016/02/013

[6] Li, M. R. (2020). Application of incident tree and hierarchical analysis method in the analysis and evaluation of mine permeability accidents. *Coal and Chemical Industry, 43*(5), 60-62, 66.

[7] Zhang, G. Q. (2020). Application of comprehensive analysis of FTA and AHP in coal mine permeability accident analysis. *Modern Mining, 36*(2), 239-241.

[8] Cheng, L. H., Xie, M. Y., Zuo, M. H., & Guo, H. M. (2022). Risk evaluation method of coal mine gas explosion based on ISM-BN and its application. *Safety in Coal Mines, 53*(10), 1-8.

[9] LEI, Y. B., Chen Z. B., Zeng, J. C., & Li, H. Y. (2016). Research on causal chain of coal mine gas accidents based on association rule. *Safety in Coal Mines, 47*(8), 240-243.

[10] Raviv, G., Fishbain, B., & Shapira, A. (2016). Analyzing risk factors in crane-related near-miss and accident reports. *Safety Science, 9*1, 192-205. https://doi.org/10.1016/j.ssci.2016.08.022

[11] Singh, K., Maiti, J., & Dhalmahapatra, K. (2019). Chain of events model for safety management: data analytics approach. *Safety Science, 118*, 568-582. https://doi.org/10.1016/j.ssci.2019.05.044

[12] Gao L. & Wu, H. (2013). Verb-Based Text mining of road crash report. *92nd Annual Meeting of the Transportation Research Board*.

[13] Nayak, R., Piyatrapoomi, N., & Weligamage, J. (2010). Application of text mining in analysing road crashes for road asset management. *Engineering Asset Lifecycle Management*, 49-58. https://doi.org/10.1007/978-0-85729-320-6_7

[14] Jia, X. F., Li, C. B., & Zhou, Y. (2023). Risk identification and influence analysis model for urban energy internet based on knowledge graph improved decision-making trial and evaluation laboratory. *Expert Systems with Applications, 233*, 120997. https://doi.org/10.1016/j.eswa.2023.120997

[15] Bai, Y. P., Wu, J. S., Ren, Q. R., Jiang, Y., & Cai, J. T. (2023). A BN-based risk assessment model of natural gas pipelines integrating knowledge graph and DEMATEL. *Process Safety and Environmental Protection, 171*, 640-654. https://doi.org/10.1016/j.psep.2023.01.060

[16] Liu, G. Y., Boyd, M., Yu, M. X., Halim, S. Z., & Quddus, N. (2021). Identifying causality and contributory factors of pipeline incidents by employing natural language processing and text mining techniques. *Process Safety and Environmental Protection, 152*(3). https://doi.org/10.1016/j.psep.2021.05.036

[17] Zelenka, M. & Podaras, A. (2021). Increasing the effectivity of business intelligence tools via amplified data knowledge. *Studies in Informatics and Control, 30*(2), 67-77. https://doi.org/10.24846/v30i2y202106

[18] Xu, G., Weng, X. L., Dan, B., & Duan, H. W., (2023). Hedging strategies of supply chain under risk aversion. *Economic Computation and Economic Cybernetics Studies and Research, 57*(1), 73-88. https://doi.org/10.24818/18423264/57.1.23.05

[19] Si, S.L., You, X. Y., Liu, H. C., & Zhang, P. (2018). DEMATEL Technique: a systematic review of the state-of-the-art literature on methodologies and applications. *Mathematical Problems In Engineering*, 1-33. https://doi.org/10.1155/2018/3696457

[20] Yazdi, M., Khan, F., Abbassi, R., & Rusli, R. (2020). Improved DEMATEL methodology for effective safety management decision-making. *Safety Science, 127*, 104705. https://doi.org/10.1016/j.ssci.2020.104705

[21] Xie, J. H., Tian, F. J., Li, X. Y., Chen, Y. Q., & Li, S. Y. (2023). A study on the influencing factors and related paths of farmer's participation in food safety governance-based on DEMATEL-ISM-MICMAC model. *Scientific Reports, 13*, 11372. https://doi.org/10.1038/s41598-023-38585-w

[22] Xiang, J. & Ma, C. S. (2023). Modeling the effectiveness of blended learning promotion with artificial intelligence adaptive learning system. *Journal of Logistics, Informatics and Service Science, 10*(3), 88-97. https://doi.org/10.33168/JLISS.2023.0307

[23] Lina, B. (2020). Import Risks for the Country: A case study in lithuania. *Journal of Service, Innovation and Sustainable Development, 1*(1&2), 51-68.

[24] Lin, Y., Hao, M. M., & Wang, Y. Y. (2022). Research on cost control of prefabricated steel structure building based on DEMATEL-ISM model. *Construction Economy, 43*(9), 54-60.

[25] Feng, S. C., Qi, C. M., Bu, B., & Chen, Y. H. (2023). Analysis on influencing factors of construction safety risk based on improved DEMATEL-ISM. *Journal of Engineering Management, 01*, 141-146.

[26] Ola, M. E. (2023). The moderating role of workplace social support in the relationship between workplace bullying and job performance. *Journal of System and Management Sciences, 13*(4), 1-15. https://doi.org/10.33168/JSMS.2023.0401

[27] Liu, S.C. (2023). Research on computational methods and

algorithms for dimensionality reduction and feature selection in high-dimensional data. *Journal of Logistics, Informatics and Service Science, 10*(3), 1-12. https://doi.org/10.33168/JLISS.2023.0301

Contact information:

Yang YANG, Professor, PhD
School of Management,
China University of Mining and Technology (Beijing),
No. 11, Ding, Xueyuan Road, Haidian District, Beijing, China
E-mail: 201329@cumtb.edu.cn

Zhilei WU, PhD
School of Management,
China University of Mining and Technology (Beijing),
No. 11, Ding, Xueyuan Road, Haidian District, Beijing, China
E-mail: 18811176258@139.com

Fenfen SHI, PhD
(Corresponding author)
School of Management,
China University of Mining and Technology (Beijing),
No. 11, Ding, Xueyuan Road, Haidian District, Beijing, China
E-mail: bqt2000503019@student.cumtb.edu.cn