Urban Rail Transit (2024) 10:65–88
https://doi.org/10.1007/s40864-023-00210-4

http://www.urt.cn/

ORIGINAL RESEARCH PAPERS

The Spatiotemporal Evolution Mechanism of Urban Rail Transit Fault Propagation in Networked Operation Modes

Ding Xiaobing[1] · **Hu Hua**[1] · **Liu Zhigang**[1] · **Mu Qingquan**[1]

Received: 22 May 2023 / Revised: 8 August 2023 / Accepted: 28 August 2023 / Published online: 25 February 2024

Abstract The cascading propagation and evolution of metro operation failures can significantly impact the safety of metro operation. To overcome this challenge, this study pre-processes a massive amount of metro operation log data through noise reduction. Moreover, a professional terminology dictionary is constructed along with a custom stop-word dictionary to segment the preprocessed data. Subsequently, the AFP-tree algorithm is employed to mine the segmented log data and identify key hazards. A weighted urban rail transit network is established, considering the effective path time cost, and the shortest travel OD path. To simulate the dynamic evolution of the failure chain propagation, a model based on disaster propagation theory is constructed. Taking the Shanghai Metro line as a case, multiple simulation scenarios are established with 25 key hazards as triggering points, and the number of cascade failure stations affected under different scenarios is outputted. The results indicate that the fault stations caused by the large passenger flow are the largest. Meanwhile, the number of stations affected by the door clamp is the smallest. The scale of fault stations reaches a maximum value in 16–20 min. Through case analysis, a positive correlation is found when the self-recovery factor is between 14 and 18, and the number of fault stations shows a significant increasing trend. The research results can provide decision-making support and theoretical guidance for rail transit operation safety management enterprises.

✉ Ding Xiaobing
dxbsuda@163.com

[1] School of Urban Rail Transportation, Shanghai University of Engineering Science, Shanghai 201620, China

Communicated by Baoming Han

Keywords Urban rail transit · Complex network · Hazard identification · Cascading failure · Fault propagation

1 Introduction

With the continuous increase in demand for rail transit in various cities, the metro network structure and scale are rapidly developing, presenting a networked and complex characteristic [1]. As the backbone of large-scale urban transportation, urban rail transit systems have two main characteristics: First, there is a strong correlation within the network—stations and lines are interconnected, and their interactions affect each other. Second, the network has limited carrying capacity: each line and node on the line has a maximum carrying capacity. Only when each node and line in the network operate within the carrying capacity range can the entire network system run smoothly. When an abnormality occurs at a node in the network, such as a sudden increase in passenger flow, natural disasters, terrorist attacks, or severe weather, a station may become overloaded or directly paralyzed. Because of the strong correlation within the urban rail transit network, faults can spread to surrounding stations, and the limited carrying capacity of stations restricts their ability to share loads. When they bear a load beyond their capacity, these stations will also become paralyzed due to overload operation, leading to a cycle of failure and even causing the entire line or network to fail, which poses great danger. Complex networks have scale-free and small-world characteristics and have been applied in multiple fields, including power facilities, communication facilities, and transportation [2, 3]. Complex networks can intuitively describe the process of node failure and find the weak links in the network. Due to the robustness of urban rail transit networks, they are not easily destroyed under

random attacks. However, when one or more stations in the network are attacked due to unexpected situations, the operating status of the stations can easily be destroyed, and the entire network may be affected by the propagation of these abnormal situations [4]. Some existed studies have evaluated the stability of the entire network by assuming the failure of a node or edge in the network.

However, these studies have inherent limitations, as they generally simulate the attack on a station in the network without considering specific risk factors that cause the station to malfunction. The impact of different types and levels of risk events on stations is also different. Additionally, the dynamic changes within the metro network over time are difficult to accurately characterize. Therefore, this paper proposes a method for predicting and estimating the cascade failure propagation trend of the subway network based on key risk sources, which can overcome these limitations.

2 Literature Review

Research on the network propagation of operational failures in rail transportation has focused primarily on exploring methods for text mining of hazard sources and failures, analyzing cascading failures, and studying the spread of disasters. This section covers these three aspects from the perspectives of text mining, cascading failure analysis, and disaster propagation.

2.1 Text Mining

In recent years, natural language processing technology has matured and has been widely applied in various fields [5–7]. In the safety field, some scholars have extracted risk item features from texts, such as Luo [8], who proposed the preprocessing of road traffic accident reports to enhance the feature representation of risk sources, build a double-hidden-layer adaptive convolutional neural network, and identify risk sources through sample training. Li [9] extracted features from reports on high-altitude construction accidents, obtained causal feature items, causal networks, and causal sets, and displayed the results using word clouds and network structure graphs. Some scholars have also used network models to mine risk factors and make predictions, such as Xue [10], who focused on the safety accident reports of construction projects, constructed a safety network model, graded the influencing factors, and implemented graded control to verify the feasibility of the model in handling such problems. Wu [11] used R language to mine ship collision accident text reports, studied methods for processing rare professional terms, and based on the mining of causal key factors, built a Bayesian network model to predict river ship accident risks. Others have implemented risk mining and control through system construction or design management frameworks. Fa [12] used coal-mining accident reports, used text mining technology to establish a coal mine human factor analysis and classification system, extracted strong association rules among influencing factors, and proposed relevant hypotheses to identify and analyze the hierarchical structure relationship in the human–machine interaction system framework from multiple perspectives. Xu [13] used text mining technology, based on data from construction accident reports, designed a translation management framework, and proposed information entropy-weighted term frequency for term importance evaluation, ultimately extracting core factors affecting construction safety. Chu [14] proposed a global supply chain risk management framework based on text mining, and collected and analyzed the existing literature; the analysis results revealed the importance of content related to terms, further defining potential supply chain risk factors. Zhao [15] proposed a network news risk factor extraction method based on the latent Dirichlet allocation (LDA) model, ultimately determining 28 risk factors, analyzing the relationships among these factors, and evaluating the risk factor structure of the oil market. Later, some scholars combined text mining technology with complex networks to find the connections between accident causation items. Qiu [16] creatively combined text mining technology with complex networks to identify 52 main accident causation factors, further constructing a coal mine accident causation network, clarifying eight core factors and their associated sets, as well as seven key links. Abdhul [17] proposed an automatic, semi-supervised, and domain-independent accident report analysis method, identified specific domain keywords in complex network structures, and grouped them into topics with expert participation, using these keywords and topics for various data mining purposes. Meanwhile, other scholars have utilized text mining techniques in practical applications to achieve quantitative analysis of accidents. For example, Liu [18] extracted train derailment accident data for various track types from the Federal Railroad Administration (FRA) Railway Equipment Accident Database and statistically analyzed them based on the frequency and severity of occurrence to derive the main causes of train derailment accidents. Wang [19] designed a railroad safety dictionary and comprehensively used algorithms including tire tree, directed acyclic graph (DAG), Viterbi, hidden Markov model (HMM) and other algorithms to extract causative keywords from accident reports, and then mined the correlation rules between the causative factors and the accidents, and combined with the high-speed railroad derailment matching model based on the risk factors of external environments to achieve an accurate and quantitative analysis of the safety situation of high-speed railroads.

The above literature focuses on the use of text mining technology to mine and analyze risks or risk causes from

accident reports in the industry, and has achieved certain results. However, this type of research generally relies on experts or focuses on the features of risk causation items for risk analysis, and there is still a need for improvement in terms of the mining of risks or risk causes from texts.

2.2 Cascading Failure

Currently, in some research on metro network failures, it is assumed that when a node or edge in the network fails, it does not affect other nodes or edges in the network. This is referred to as static robustness research. However, in complex networks in the real world, such as urban rail transit networks, power networks, and communication networks, some nodes or edges may fail due to random accidents or deliberate attacks, causing cascading failures that may affect other nodes or edges in the network, leading to a chain reaction; this phenomenon is known as cascading failure [20]. Experts and scholars in various fields have extensively studied the cascading failure process, and the models proposed for cascading failure mainly include the load-capacity model [21], binary model [22], and sandpile model [23]. Among them, the load-capacity model has the widest impact and has been applied in empirical research and analysis of real networks.

Research on the load-capacity model focuses mainly on three basic issues: the definition of the initial load of nodes or edges in the network, the definition of the capacity of nodes or edges, and the method of load redistribution. Freeman [24] defined the initial load of nodes as their betweenness centrality, and the capacity of nodes as a linear function of the initial load, which is a reasonable and widely used characterization. However, considering that betweenness centrality is a global quantity and the calculation complexity is high, it is necessary to obtain the global properties of the entire network. Later, Wang [25] and others defined the initial load of nodes based on the degree of nodes and the total degree of adjacent nodes, which proposed a new concept of the probability of overload node failure.

Currently, commonly used load redistribution methods can be divided into two categories: one is based on the global allocation of the entire network, and the other considers the nearest allocation strategy of the capacity of adjacent nodes to failed nodes. Li [26] posited that the information processing capacity of nodes could be reflected by the size of node degree, and effective allocation of additional capacity through vertex quotas could prevent cascading failure and effectively improve the robustness of the network. Duan [27] proposed a cascading failure model with adjustable load redistribution range and load redistribution heterogeneity, and analyzed the cascading failure conditions of the model on scale-free networks. The results showed that reasonable adjustment of the load redistribution range and heterogeneity can significantly improve the robustness of complex networks. Fang [28] introduced the concept of neighbor links and proposed a load distribution method which can average the load of failed nodes to their adjacent nodes, and studied the cascading failure phenomenon on directed complex networks in a new environment. Ma [29] proposed a new load-capacity model, which redefined the load distribution rule based on the self-repair time factor of nodes and analyzed the adjustable parameters of node capacity and self-repair factor. Ju [30] combined the degree and betweenness centrality of nodes, and redefined the load distribution of adjacent edges, studying the robustness of network cascading failure. Li [31] constructed a model of urban passenger transport network in the city cluster, and evaluated the anti-destructive performance of cascading failure by adopting an improved optimal load allocation strategy based on actual passenger flow weighting.

Analyzing and summarizing the above domestic and foreign research, it can be found that most research on cascading failure in metro networks is based on real cities, and through the analysis of the network structure and the study of cascading failure models, it provides a theoretical basis and decision-making support.

2.3 Disaster Propagation

Buzna et al. [32] first proposed a model of fault propagation in a general network system that considers node recovery capability and transfer mechanisms to describe the dynamic spread and impact of disasters in complex networks. The model combines network nodes into active bistable elements with delayed interactions along directed links. Later, Buzna et al. [33] applied disaster propagation theory to study the effectiveness of different emergency strategies and optimized resource allocation based on network state and topology. By changing network topology, delay time factors, and overall resource allocation, the effectiveness of different emergency strategies was evaluated. Hu [34] proposed a resource node attribute model based on disaster propagation theory, which combines resource value, disaster energy of each node, disaster propagation path, and disaster propagation characteristics. Finally, the optimal timing for disaster relief and emergency resource preparation was determined through the model. Yi [35] used a method for simulating multiple failure events to describe the random factors that trigger disasters. Ouyang [36] presented an improved model of redundant systems in networks, and analyzed the differences in the spreading process and the role of important parameters. The results show that the disaster spreading process becomes slower with the existence of redundant systems in the network. Later, Ouyang [37] studied the impact of several redundancy strategies on controlling disaster propagation in a network and found that an improved random network can better cope with disasters, while the strategy based on total degree is the

most effective way to control disaster propagation in scale-free networks. Weng [38] established a universal disaster propagation dynamic model and studied the influence of three important characteristic parameters: self-repair factor, delay time factor, and noise intensity. Xiao [39] established a dynamic model of congestion propagation in the rail transit network based on the disaster propagation dynamic model. The congestion propagation model can reflect the process of congestion propagation in the rail transit network, and the simulation process reveals the propagation law of congestion in the metro network.

In summary, the early research on disaster propagation theory fully considered the evolution process of faults over time, node self-recovery capability, disaster-fault-attack propagation mechanisms, and other influencing factors such as internal random noise. Based on disaster propagation theory, combined with the characteristics of the subway network and key hazards, a cascade failure model can be established to explore the propagation mechanism of faults in the network when it suffers from different forms and levels of attacks. Data can be used to predict the scope and degree of different risks.

3 Identification and Quantitative Treatment of Key Hazard Sources

3.1 Methods for Identifying Key Hazard Sources

Currently, there is still a problem of unknown high-risk sources in the operation process of subway systems. In the actual operation process, identification mainly relies on experienced experts or staff, which has a high degree of subjectivity and lacks scientific and effective data support. Therefore, establishing a key hazard identification method based on subway dispatch logs is a new exploration, which can accurately identify key hazards from a data perspective.

3.1.1 Data Preprocessing

The dispatch logs contain a large amount of information, but they typically consist of textual descriptions of events that cannot be directly used as objects for data mining. Therefore, data preprocessing is required, and a data processing flow as shown in Fig. 1 is designed to handle the data.

The first step is to clean the interfering data. Since the dispatch logs contain a large number of records that are irrelevant to the operation risk events, including normal vehicle dispatch and routine maintenance information, which do not contain risk source information, Python language is used in combination with common hazard sources to filter the data, and extract most of the valid data, as shown in the pseudocode in Fig. 2. Meanwhile, to ensure the integrity of the valid data, other data are manually screened to achieve completely cleaning of the interfering data.

Step 2: Word segmentation and stop-word removal. Using the Pycharm development platform and the Jieba library for word segmentation in Python, a more suitable, accurate mode for text analysis was used. A custom professional terminology dictionary was loaded, and then it was segmented to improve the accuracy of the segmentation. After segmentation, stop words were removed to delete words in the log text that are not relevant to the research, such as "punctuation", "numbers", "also", "just" and other words, based on the Harbin Institute of Technology stop-word list and customized stop-word dictionary, taking into account the professional characteristics of subway operation. The specific processing results are shown in Table 1.

Step 3: Data format conversion. Assuming that after preprocessing the log file, each inputting data contains n words, then two consecutive inputting data can be represented as shown in Eqs. (1) and (2):

$$A_1 = (a_1, a_2, a_3, ..., a_n) \tag{1}$$

$$A_i = (a_n, a_{n+1}, a_{n+2}, ..., a_j) \tag{2}$$

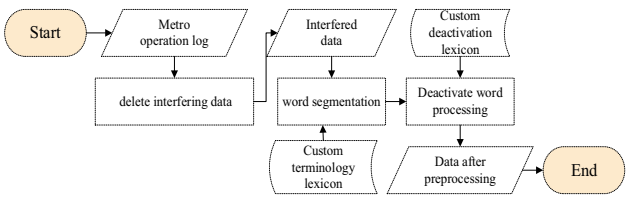

Fig. 1 Data preprocessing process

```
import pandas as pd
data = pd.read_excel(r'data.xlsx', 'r')
List = [ ]
for i in sentence:
if a = (‘Normal scheduling’):
   del a
else
   list.append(a)
if b =(‘Routine maintenance’):
     del b
   else
   list.append(b)
……
return list
# i is the iteration variable
```

Fig. 2 Pseudocode removal of jamming data

Table 1 Jieba word segmentation and removing stop words

Raw data	Jieba word segmentation	Stop words removed
At 14:33, the driver of train 914, car 0908 on the downbound JiaShan Road reported a train broadcasting malfunction. Maintenance was notified, but due to the driver handling an object caught in the door at YiShan Road station and an extended station stop, the train arrived at the terminal station 5 minutes late, causing the following 2 trains to also be delayed by 5 minutes.	14/:/33/Jiashan Road/down/914/time/0908/#/train/driver/report/train/broadcast failure/,/notice/maintenance/parking/,/due to/in/Yishan Road/platform/door clamping/driver/handling/and/stop extension/,/the train/final arrival/delay/5/minute/,/and/cause/follow-up/2/train/delay/5/minute	Jiashan Road/down/driver/report/train/broadcast failure/notice/maintenance/attendance/due to/Yishan Road/platform/door clamping/driver/handling/and/stop extension/this train/final arrival/delay/cause/follow-up/train/delay/minute/
At 21:15, the driver of train 15096, car 0103 on the downbound Lianhua Road reported a malfunction of the first door of the third car. After multiple attempts to open and close the door, it was ineffective. The train control dispatched the train to the site for door removal, and the train resumed normal operation after the door was removed. The malfunctioning train caused a delay of 2 minutes at the station, but arrived at the terminal station on time.	21/:/15/Lianhua Road/down/15096/time/0103/#/car/driver/report/section 3/first/door/fault/passing/multiple times/opening/closing/door/operation/after/invalid/dispatching/order/its/arrival/site/door removal/after/recovery/normal operation// Fault/train/cause/this station/late departure/2/minute/,/final arrival/late arrival//	Lianhua Road/Down/Train/Driver/Report/Section 3/First/Door/Fault/Pass/Multiple/Switch/Door/Operation/After/Invalid/Traffic Control/Order/Site/Door Removal/After/Resume/Normal Operation/Fault/Train/Cause/This Station/Late Departure/Minute/Final Arrival/Not Late/

Table 2 Event description format after preprocessing of operation log

Serial number	Operational event description	Data storage format
01	Zhaojiabang/up/section 2/the third door/door/fault/cause/the train/final arrival/delay	$a_1/a_2/a_3/a/a/a_6/a_7/a_8/a_9/a_{10}$
02	Yishan Road/up/third car/carriage/whole car/door/fault/final arrival/delay/minute	$a_{11}/a_2/a_{12}/a_{13}/a_{14}/a_5/a_6/a_9/a_{10}/a_{15}$
03	Shanxi South Road/up/section 5/second door/door/unable/closed/notice/follow-up/station/estimated/delayed/minute	$a_{16}/a_2/a_{17}/a_{18}/a_5/a_{19}/a_{20}/a_{21}/a_{22}/a_{23}/a_{24}/a_{10}/a_{25}$

in which A_i represents the ith data entry, $i \in (1, m)$, a_j represents the embedding of the jth word, $j \in (1, n)$, for the same word, using the same a_j to represent it. Similarly, all words can be represented in the form of a_j.

The final storage format for the log data is shown in Table 2.

The serial number in Table 2 represents the independent code of each log data.

3.1.2 Algorithm for Identifying Key Hazards

The classic Apriori algorithm first generates candidate items and their corresponding support through connection, and then filters out frequent item sets based on a support threshold. This algorithm achieves feasible association rule extraction on large datasets, but it requires multiple repeated accesses to the transaction database during the calculation process. When analyzing large amounts of data, this can lead to excessive I/O load, and the calculation process can result in excessively large candidate sets, ultimately leading to insufficient computer memory and greatly increased time costs. To overcome these shortcomings, Han [40] proposed the frequent pattern (FP)-growth algorithm, which integrates the scanning of elements in the database into an FP-tree and still retains the association information in the item set. This algorithm can complete the analysis with only two passes of the database, but it still requires a long time to process large datasets and occupies a large amount of computer memory, resulting in poor computational efficiency. Drawing on the processing ideas of the FP-growth algorithm, a more efficient association rule algorithm was designed by improving its analysis efficiency.

In the calculation process, the two indicators of support and confidence are used as the basis for judgment in the data processing process. The support is used to calculate the probability of the occurrence of data associations, while the confidence is used to mine strong association rules in the text. The formulas for calculating the two indicators are shown in Eqs. (3) and (4).

$$Support(X, Y) = P(XY) = \frac{m_{XY}}{M_{all}} \tag{3}$$

$$Confidence(X \Leftarrow Y) = P(X|Y) = \frac{P(XY)}{P(Y)} \quad (4)$$

where X and Y respectively represent different data elements, $P(XY)$ represents the probability of X and Y occurring simultaneously, m_{XY} indicates the frequency of X and Y occurring at the same time, and M_{all} represents the total amount of data.

$P(X|Y)$ represents the probability of X occurring under the condition of Y occurrence, $P(Y)$ represents the probability of occurrence of element Y.

The first step of the algorithm is to build the FP-tree. First, the entire dataset is scanned to accumulate and count the frequency of all items. Then, the data that do not meet the support threshold according to the set support are filtered out, and the remaining are sorted data from high to low to generate a table of frequent item set events. In this study, the setting of the support threshold needs to be determined based on the actual data and research direction. The support count can be temporarily set to 2 for the purpose of illustration. Therefore, as shown in Table 3, the re-sorted data such as a_{16}, a_{17}, a_{18} can be removed.

After determining the frequent item set events, a header table must be built to store the occurrence frequency of all item sets. In this condition, the pointer points to the first node of the corresponding item in the tree. In Python programs, the "dictionary" is used to store the header table, and the final constructed result is shown in Fig. 3.

After constructing the FP-tree, frequent patterns need to be mined from the tree. However, when generating conditional pattern bases in the FP-growth algorithm, multiple traversals of common paths are required. When constructing a large FP-tree, the algorithm will occupy a large amount of computer memory and significantly prolong computation time [39, 40]. To improve algorithm efficiency and reduce situations where the algorithm complexity is too high when computing large amounts of data, the ascending FP (AFP)-tree algorithm was proposed as an improvement to the FP-growth algorithm. The AFP-tree algorithm utilizes the preorder traversal concept to read the FP-tree, such that obtaining all conditional pattern bases for frequent one-item sets only requires scanning the FP-tree once. The basic steps for generating conditional pattern bases using the AFP-tree algorithm are as follows:

(1) Build a common path (CP) with the initial value set to null. Scan node a_5, and the CP stores the prefix path of a_5. Since CP is currently empty, the conditional pattern base of a_5 is also empty.
(2) Add a_5 to CP, then scan node a_{10}. At this point, SP stores the prefix path of a_{10}. Therefore, a_5 is a conditional pattern base of a_{10}, with a support count of 4, denoted as a_5:4.
(3) After storing a_{10} in CP, update the content of CP as a_5a_{10}. Then, scan a_2. At this time, CP stores the prefix path of a_2, so a_5 and a_{10} are the conditional pattern bases of a_2 with a support count of 3, denoted as a_5,a_{10} :3.
(4) Store a_2 in CP and update CP to $a_5a_{10}a_2$. Then scan a_6, and following the same process as in steps (2) and (3), we can obtain the prefix path for a_6. Continuing the scan, we obtain the prefix path for a_8 as a_5,a_{10},a_2,a_6,a_9 ,a_4,a_7:1. At this point, we realize that a_8 is a terminal node, so return to the most recent branching node and traverse the unexplored branch node a_9. At the same time, update CP to $a_5a_{10}a_2a_6$.

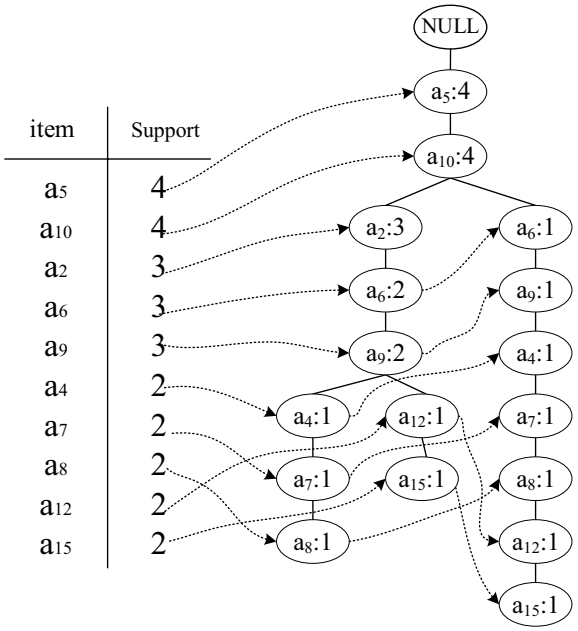

Fig. 3 Construction of FP-tree

Table 3 Frequent item set event element table

Number	Elements of the initial event description	Elements after filtering and reordering
01	$a_1/a_2/a_3/a_4/a_5/a_6/a_7/a_8/a_9/a_{10}$	$a_5a_{10}a_2a_6a_9a_4a_7a_8$
02	$a_{11}/a_2/a_{12}/a_{13}/a_{14}/a_5/a_6/a_9/a_{10}/a_{15}$	$a_5a_{10}a_2a_6a_9a_{12}a_{15}$
03	$a_{16}/a_2/a_{17}/a_{18}/a_5/a_{19}/a_{20}/a_{21}/a_{22}/a_{23}/a_{24}/a_{10}/a_{25}$	$a_5a_{10}a_2$
04	$a_{26}/a_{27}/a_{12}/a_4/a_5/a_6/a_7/a_8/a_9/a_{10}/a_{15}$	$a_5a_{10}a_6a_9a_4a_7a_8a_{12}a_{15}$

Table 4 Finding frequent patterns through conditional pattern bases

Item	Conditional pattern base	Generated frequent patterns
a_5	(∅ :4)	–
a_4	(a_5,a_{10},a_2,a_6,a_9:1), (a_5,a_{10},a_6,a_9:1)	(a_5,a_4:2), (a_5,a_{10},a_4:2)
a_9	(a_5,a_{10},a_2,a_6 :2)	(a_5,a_{10},a_2,a_6,a_9:2)
a_6	(a_5,a_{10},a_2 :2), (a_5,a_{10} :1)	(a_5,a_{10},a_6:3)
a_{10}	(a_5 :4)	(a_5,a_{10} :4)

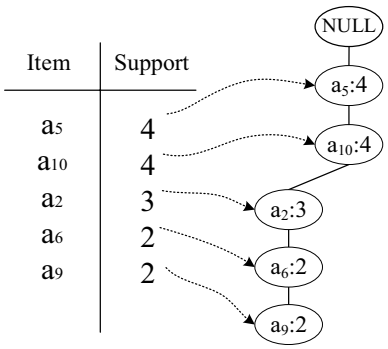

Fig. 4 Conditional FP-tree

(5) Continue scanning the other child node a_{12} of x_9, obtaining a conditional pattern base of a_5,a_{10},a_2,a_6,a_9:1 for a_{12}, and updating CP content to $a_5a_{10}a_2a_6a_9a_{12}$. Then scan a_{15}, obtaining a conditional pattern base of a_5,a_{10},a_2,a_6,a_9,a_{12}:1 for a_{15}.
(6) Continuing the scan, it is found that a_{15} is a leaf node, and then returns to the unscanned branch node a_{10}. By repeating this process, all remaining subnodes in the tree are scanned and all conditional pattern bases are obtained, as shown in Table 4.

The AFP-tree algorithm scans the tree using the idea of preorder traversal; only one scan of all nodes in the FP-tree is needed to obtain the conditional pattern bases of all frequent one-item sets in the data. The complexity of the algorithm, including time and space complexity, is the same as the number of nodes in the tree, which is O(n), where n is the total number of nodes in the FP-tree.

Furthermore, real-time pruning is used to sort the frequent item sets in descending order of their support and only keep those that meet the support threshold, deleting the items that do not meet the threshold. This results in a non-redundant conditional FP-tree as shown in Fig. 4.

From the mining result in Fig. 4, we can obtain a frequent pattern: $(a_5, a_{10}, a_2, a_6, a_9, : 2)$, which indicates a strong association between "door", "delay", "up", "fault", and "terminal". Furthermore, we can derive (a_5,a_{10} :4), indicating an even stronger association between "door" and "delay". Because of the limited sample size in this example, it is not possible to fully describe all the association rules that may exist in the log data. However, with a sufficiently large sample size, more frequent item sets can be discovered, and the complete rules can be derived through text mining, allowing us to identify the key risk factors that cause operation risks in the subway system.

3.2 Weighted Identification of Key Risk Sources

3.2.1 Sequential Relationship Weighting Analysis

The subjective weighting method is commonly used in sequence relationship weighting analysis, which includes the analytic hierarchy process (AHP) and sequence relationship analysis. AHP, as the most widely used subjective weighting method for considering weight issues, increases the complexity of expert judgment on pairwise factors when judging multiple-factor indicators, which can easily cause logical confusion and increase the difficulty of consistency judgment and accuracy. Therefore, AHP is not suitable for the research content of this problem. Sequence relationship analysis effectively avoids the logical problems and huge workload caused by the large number of factors, mainly by comparing each key hazard source with the others through the experience and cognition of relevant domain experts, ranking the relative importance of each key hazard source, and determining their subjective weight. The idea of using expert experience to rank relative importance is consistent with the idea of identifying key hazard sources by mining and determining them, and the data from the previous analysis can be used to fill in the part that requires expert judgment, further reducing the impact of subjective factors on the weighting analysis results. Therefore, the sequence relationship analysis method is adopted for subjective weighting of key hazard sources.

The main steps of the sequence relationship analysis method are as follows:

(1) Determining the importance of indicators and establish their sequence relationships.

Invite relevant experts to rank the relative importance of key hazard sources, if the relative importance of hazard source H_i is higher than that of hazard source H_j, it is denoted as $H_i > H_j$.

Sort each hazard source in order of relative importance as $H_a > H_b > ... > H_m > H_n(a, b, m = 1, 2, ..., n)(2)$ C a l culating the relative importance of adjacent hazardous sources.

Table 5 Values for relative importance of hazard sources

Order number	Importance	Explanation of H_i compares with H_j
a	1.0	Equally important
b	1.1	Slightly more important
c	1.2	More important
d	1.3	Between important and significantly important
e	1.4	Significantly more important
f	1.5	Moderately more important
g	1.6	Strongly more important
h	1.7	Between strongly important and extremely important
i	1.8	Extremely important

Once the relative importance of each key hazard source is determined through expert judgement, the relative importance of each adjacent hazard source needs to be determined. The relative importance R_k between hazard source H_k and its adjacent source H_{k-1} can be expressed as Eq. (5), and the values for relative importance R_k are given in Table 5 [43].

$$R_{\mathrm{k}} = \frac{H_k}{H_{k-1}}, k = n, n-1, ..., 3, 2 \tag{5}$$

If the product of the relative importance values $\Pi_{k=1}^{n} R_k > 1.8$, the cumulative importance degree judged is greater than the extreme importance degree, indicating an abnormality in the subjective judgment. It is necessary to correct R_k by the correction coefficient μ according to Eqs. (6) and (7), and the values for relative importance of hazard sources are shown in Table 5.

$$\mu = \left(\frac{1.8}{\Pi_{k=1}^{n} R_k}\right)^{\frac{1}{n-1}} \tag{6}$$

$$R_k' = \mu \cdot R_k \tag{7}$$

(3) Calculating subjective weight values

According to the relative importance degree of risk factors, the weight of key hazard sources is calculated as Eqs. (8) and (9).

$$\omega_n' = \left\{ \begin{array}{l} (1 + \sum_{\mathrm{i}=1}^{n} \prod_{k=i}^{n} R_k)^{-1}, \prod_{k=i}^{n} R_k \le 1.8 \\ (1 + \sum_{\mathrm{i}=1}^{n} \prod_{k=i}^{n} R_k')^{-1}, \prod_{k=i}^{n} R_k > 1.8 \end{array} \right\} \tag{8}$$

$$\omega_{k-1}' = R_k \omega_k' \quad (k = n,\ n-1, \ldots, 2) \tag{9}$$

Calculate the weight of ω_k', and obtain the weight set as $\omega_k' = (\omega_1', \omega_2', ..., \omega_n')$ in the end.

3.2.2 *Objective Weighting Analysis*

Currently, the mainstream objective weighting methods in academic research include the entropy method, critic method, and principal component analysis [44]. These methods are based on the characteristics of the indicator values themselves to weight them, and have demonstrated good normative properties but still require experts for evaluation and scoring before weighting.

Therefore, combined with the characteristics of key hazard sources, the entropy method was selected for objective weighting. Since the confidence value is calculated based on the association rules in the data, its idea is consistent with the judgment of key hazard sources by experts when scoring. The difference is that the confidence value is calculated objectively through data analysis, while expert scoring is obtained based on the expert's subjective wishes. Therefore, replacing the original expert scoring values with confidence values can make the weighting results more objective. Since subway transportation passenger flow is affected to a certain extent by seasons, the original data are divided into four categories according to seasons. The entropy method is used to determine the objective weighting values, and the specific calculation steps are as follows:

Step1: Determine the initial values of each hazard source indicator and construct the initial data matrix.
Step2: Standardize the initial data; suppose there are k key risk sources given as $A_1, A_2, A_3, ..., A_k$, and A_k is calculated as Eq. (10):

$$A_k = \{a_{k1}, a_{k2}, a_{k3}, a_{k4}\} \tag{10}$$

Standardize the data of each key hazard source, and the standardization process Y_{kl} is shown in Eq. (11):

$$Y_{kl} = \frac{a_{kl} - \min(a_k)}{\max(a_k) - \min(a_k)} \tag{11}$$

Step3: Calculate the information entropy of each key hazard source. According to the definition of information entropy in information theory, the information entropy E_l of a set of data can be determined by Eq. (12).

$$E_l = -\frac{1}{\ln k} \sum_{k=1}^{n} p_{kl} \ln p_{kl} \tag{12}$$

where the calculation of p_{kl} is as shown in Eq. (13):

$$p_{kl} = Y_{kl} / \sum_{k=1}^{n} Y_{kl} \tag{13}$$

if $p_{kl} = 0$, then $\lim_{p_{kl} \to 0} p_{kl} \ln p_{kl} = 0$

Step 4: Determine the weights of each hazard source. According to the formula for calculating information entropy, the information entropy of each hazard source can be calculated as $E_1, E_2, E_3, ..., E_K$. The weights of each key hazard source are calculated through the information entropy as shown in Eq. (14):

$$w_l'' = \frac{1 - E_k}{k - \sum E_k} (k = 1, 2, ..., n) \tag{14}$$

3.2.3 Composite Weighting Method

In the process of weighting the key hazards, since the position and role of each hazard in causing the hazardous event are different, the accuracy of the calculation results of the constructed model is determined by the weights of different hazards. In order to better reflect the impact of key hazards on the operation of the subway network and to avoid subjective arbitrariness of the results to the maximum extent, the subjective weights and objective weights calculated above are combined using the multiplication and addition method to calculate the comprehensive weight, as shown in Eq. (15):

$$w_l = \frac{w_l' \times w_l''}{\sum_{l=1}^{n} w_l' \times w_l''} \tag{15}$$

4 Construction of a Network Failure Propagation Model Based on Key Hazard Sources

Complex networks can effectively analyze the propagation mechanism and process of hazard sources in networks, and are currently widely used for analysis and traffic flow allocation in metro networks.

4.1 Construction of a Complex Network Based on Subway Network Topology Structure

The subway undertakes a large number of urban passenger transport tasks. Similar to urban road traffic networks and railway networks, the subway network also has some common characteristics, including (1) a large number of nodes with close connections between nodes, (2) the connection structure in the network reflecting obvious self-organizing rules, and (3) various forms of local area networks, such as radial and grid-shaped. At the same time, the subway network has its own characteristics, which significantly affect the failure chain in the network: (1) The hierarchical structure—the subway transportation network generally includes three hierarchical structures: infrastructure, trains, and passenger flow. (2) Dynamic imbalance of passenger flow—the dynamic imbalance of passenger flow in the subway transportation network is mainly reflected in the obvious time and spatial imbalances of traffic demand, such as suburban and bustling commercial areas, and peak and off-peak periods. (3) Nonlinear correlation—there is a nonlinear correlation and influence between infrastructure, trains, and passenger flow in metro operation network. Combined with the theory of disaster propagation, the occurrence and development of risk events in a complex network will be transmitted and amplified by the mutual coupling existing in the network topology structure, which leads to periodic fluctuations in the stability of the entire network.

The subway topology network consists of stations and connecting lines. The Space-L and Space-P methods are the two most commonly used methods for constructing complex networks. After a comprehensive comparison of the two methods, the Space-L method is better able to express the actual topological structure of the subway network in geographical space. Therefore, the Space-L method is adopted to construct the metro topological structure. Due to the significant impact of passenger transfers and train reversals on the topology network, these nodes must be handled separately. The number of transfers and transfer time have a significant impact on the selection of passenger travel paths. Therefore, transfer stations are converted into two nodes of different lines and connected, and the passenger transfer time is used as the weight to assign the connection edge. At this time, the transfer station has the same name but a different number in the network, as shown in S1 and S2 in Fig. 5. When a station has a serious failure that causes the train to be unable to stop or operate at that station, a turnaround station can be set up to reduce the impact of a station failure on the entire network. At the same time, the topological structure of the subway network needs to be updated when a station failure occurs, as shown in Fig. 6.

In Fig. 6, when the normal station S4 malfunctions, normal station S5 cannot operate normally due to the lack of a turning point, while other stations on the line can still operate normally. When the turning point station S3

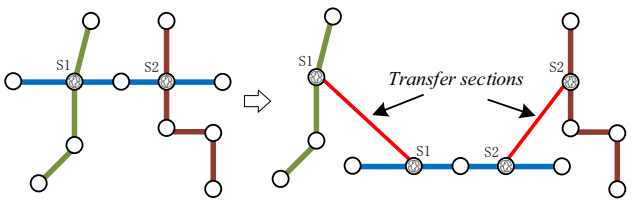

Fig. 5 Handling of transfer stations in topological network

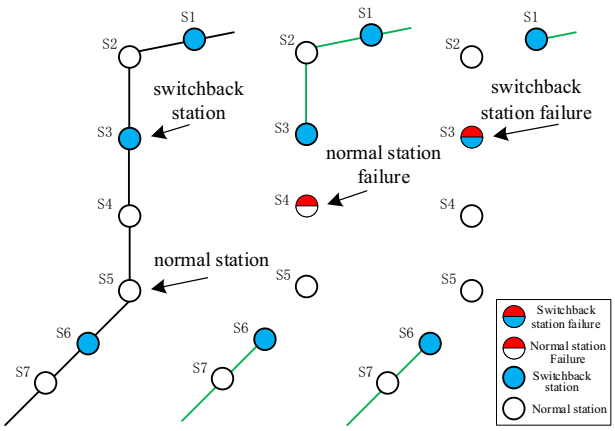

Fig. 6 Common station and turn-back station processing under fault condition

malfunctions, normal stations S2, S4, and S5 also cannot operate normally. In the operation of the subway system, the actual conditions of different stations and line sections are different, and it is necessary to weight each part of the subway network to construct a weighted network as shown in Eq. (16).

$$N_w = \{S_i, E_{ij}, W\}$$
$$\begin{cases} S_i\ i \in [1,n] \\ E_{ij}\ i,j \in [1,n]\ i \neq j \\ W = \{W_{s_i}(t), W_{\overline{s_i}}(t), W_{e_{ij}}(t)\} \end{cases} \tag{16}$$

where S_i is a set of nodes,E_{ij} represents the edge that connects node s_i and node s_j. W is the set of weights of the nodes and edges, $W_{s_i}(t)$ is the weights of nodes, $W_{\overline{s_i}}(t)$ is the transferring weight of exchange station $\overline{s_i}$,$W_{e_{ij}}(t)$ is the weight of each node, and t is time. The degree of node s_i can be expressed as $D_i = \sum x_{i\_j}$,x_{i_j} is the connection between node s_i and node s_j. If node s_i can be connected with s_j,$x_{i\_j} = 1$; or $x_{i\_j} = 0$.

The weight W needs to be determined based on the characteristics of different parts within the subway network, $W_{e_{ij}}(t)$ is determined by the train running time on the line between corresponding stations i and j, including the running time and stopping time of the train, which are fixed times that can be determined by the train timetable. $W_{s_i}(t)$ is determined by the time it takes for passengers to enter and exit station s_i.$W_{\overline{s_i}}(t)$ is mainly determined by the passenger transfer time at the transfer station $\overline{s_i}$. In the event of station failures, there may be a large number of stranded passengers. Therefore, the passenger transfer time can be calculated taking into account the efficiency of escalators inside the station, as shown in Eq. (17).

$$W = \{W_{s_i}(t), W_{\overline{s_i}}(t), W_{e_{ij}}(t)\}$$
$$\begin{cases} W_{s_i}(t) = T_i' + T_i'' + \rho_i[K_i(t)/O_t]^{\sigma_i}, \forall i \\ W_{\overline{s}_i}(t) = S_i/V_i + 1/2F_a + [E_i(t)/2n_i\mu_i - S_i/4n_i\mathrm{v}_i] \\ W_{e_{ij}}(t) = T_{e_{ij}} \end{cases} \tag{17}$$

T_i' represents the time that passengers spend entering the station, and T_i'' represents the time that passengers spend exiting the station. $\rho_i[K_i(t)/O_t]^{\sigma_i}$ represents the travel delay time caused by station malfunctions for passengers.

$K_i(t)$ represents the passenger flow at station s_i at time t, and O_t represents the maximum capacity of the train. σ_i and ρ_i are two congestion delay parameters, S_i/V_i is the transfer time, S_i is the transferring distance, V_i represents the average speed of passengers transferring. F_a represents the departure interval of the trains. $\frac{1}{2}F_a$ represents the average waiting time of passengers.$[E_i(t)/2n_i\mu_i - S_i/4n_iV_i]$ represents the transfer congestion delay time, $E_i(t)$ is the maximum queuing capacity, n_i represents the number of various escalators in the station, μ_i is the output rate of automatic and pedestrian escalators, and $T_{e_{ij}}$ represents the train's running time on a certain section of the line.

Overall, when calculating the weight of passenger travel paths, the time cost of each effective path is used as the weight to construct the calculation model, as shown in Eq. (18):

$$W_{od} = \sum_{i=1}^{n} w_{e_{ij}}(t) + w_{s_i}(t) + w_{\overline{s_i}}(t) \tag{18}$$

To calculate the shortest travel path, the occurrence of faults can cause the failure of certain stations and the spread of faults in the network can change the topology of the entire subway network. Therefore, the travel cost of a certain route may also change over time. In addition, the failure of a node does not mean that every affected node cannot operate normally, so the calculation of the shortest path should consider the influence of passenger flow changes on weight calculation. By calculating the cost of the shortest travel time, the shortest path for an OD can be obtained. After determining the faulty station and the type of fault, each cycle of the cascading fault calculation involves updating the network topology, calculating the shortest path, changing passenger flow, changing weights, and changing node status. The flow chart is shown in Fig. 7.

The Dijkstra algorithm is chosen to determine and process the OD considering congestion changes and cascading failures. This algorithm starts from the origin node and gradually expands outward to find the shortest path, with the destination node as the expansion endpoint. The network's node set is divided into two parts: S is the set of nodes where the shortest path has been found, and U

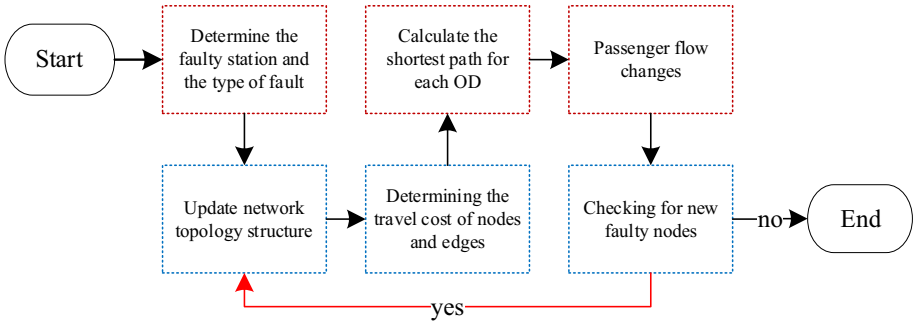

Fig. 7 Flowchart of shortest path calculation.

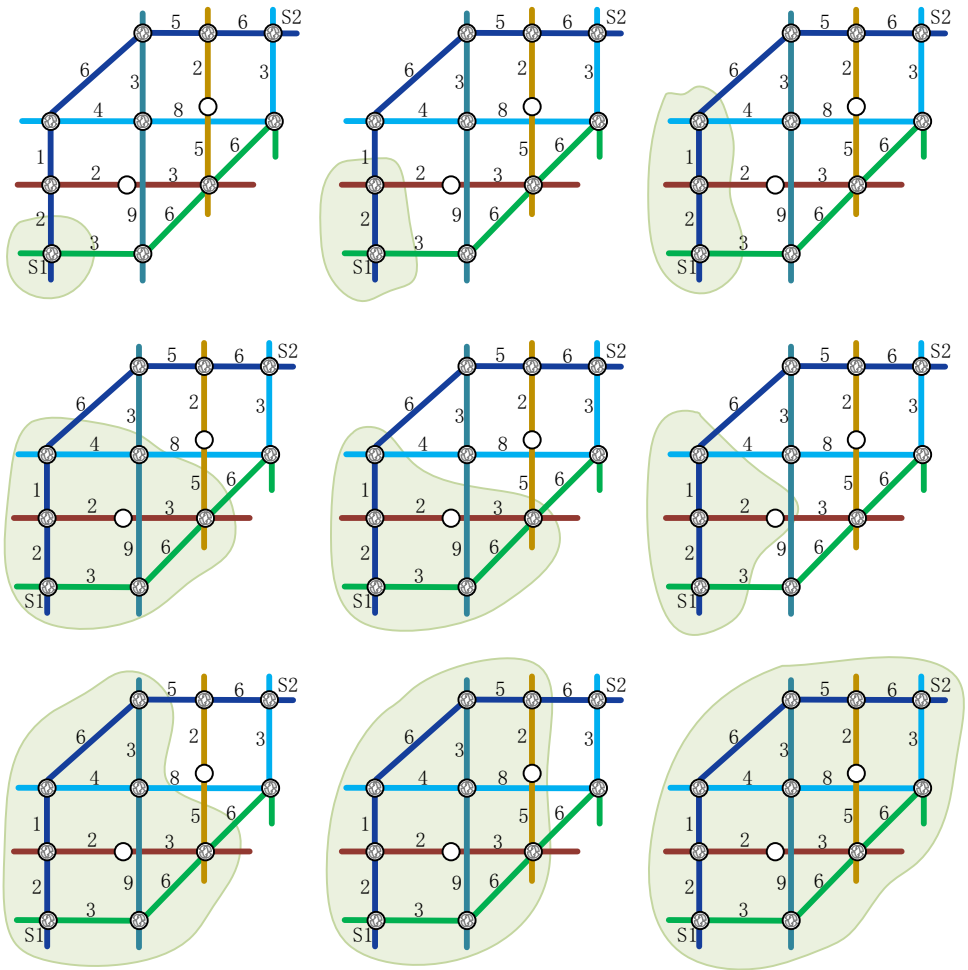

Fig. 8 Schematic diagram of the Dijkstra algorithm for calculating the shortest path

is the set of remaining nodes where the shortest path has not yet been found. When a new shortest path is found, the nodes on the path should be added to the set S until all nodes in U are added to S. The calculation process is shown in Fig. 8.

4.2 Construction of a Network Chain Fault Propagation Model

The network fault chain propagation model for a metro system is established based on the theory of disaster propagation. The following three assumptions are made when applying this theory:

(1) Two node states exist in the subway network: a normal state and a fault state caused by internal events. The nor-

mal state refers to the normal operation of a node, and the fault state refers to the state when a normal node becomes a faulty node due to an internal risk event, or when the state of a nearby node changes due to the propagation of faults in other stations. When a fault node appears, the carrying capacity of that node may be reduced or even lost completely.
(2) The occurrence of chain faults is not only affected by internal factors of the nodes, but also by random external environmental interference.
(3) The station states will discretely change over time and have a certain self-recovery capability, that is, the stations themselves have a certain robustness.

According to the theory of disaster propagation [32], the attribute value of node S_i is defined as $x_i(t), i \in \{1, 2, 3, ..., n\}$. The time evolution dynamics of attribute value $x_i(t)$ under the combined effects of the self-recovery mechanism, disaster propagation mechanism, and other factors should satisfy Eq. (19):

$$\frac{\mathrm{d}x_i(t)}{\mathrm{d}t} = -\frac{x_i(t)}{\tau_i(t)} + \Theta_i[x_i(t)]\left\{\sum_{i\neq j}\frac{M_{ij}(t)x_j[t-T_{ij}(t)]}{f[O_i(t)]}\right\} \times \exp\left[\frac{-\beta T_{ij}(t)}{\tau_i(t)}\right] + \zeta_i(t), \quad i,j \in 1,2,3,...,N, i \neq j \tag{19}$$

The left-hand side of the equation represents the trend of the node attribute value under the combined effects of the three mechanisms. The right-hand side of the equation represents the sum of the three mechanisms, the detailed explanation is provided below:

(1) Node attribute value.

The attribute value $x_i(t)$ qualitatively describes the state of node S_i in the network,$x_i(t) \in [0, 1], \forall i[0, 1], \forall i$. When $x_i(t) = 0$, the node is in a normal and stable state, when $0 < x_i(t) < 1$, the node is in an unstable and volatile state. The larger the value of $x_i(t)$, the more unstable the node is. When $x_i(t) = 1$, the node experiences a fault. Therefore, when studying network chain faults, the weighted key hazard source attribute value in Eq. (15) is used to represent the initial attribute value of the node. $\varphi_i(t)$ is used to describe the initial attribute value of the node, as shown in Eq. (20):

$$\varphi_i(t) = \left\{w_l^1, w_l^2, w_l^3, ..., w_l^n,\right\} \tag{20}$$

The sigmoid function [32] is used to define $x_i(t)$ in equation, as shown in Eq. (21):

$$x_i(t) = \begin{cases} 0, \varphi_i(t) \leq \theta_i \\ \dfrac{1}{1+\lambda\exp\left\{-\frac{[\varphi_i(t)-\theta_i]\delta}{1-\theta_i}\right\}}, \text{otherwise} \end{cases} \tag{21}$$

in which$\lambda = 15,\delta = 5$ [43], θ_i is the threshold value for a station to withstand the fault state.

(2) The self-recovery mechanism of nodes

$-x_i(t)/\tau_i(t)$ represents the self-recovery mechanism of node S_i, which shows the self-recovery mechanism of the nodes. The self-recovery factor $\tau_i(t)$ describes the node's ability to resist disasters and maintain stability, while $1/\tau_i(t)$ can be viewed as the node's self-recovery capacity and speed. As $\tau_i(t)$ increases, $-x_i(t)/\tau_i(t)$ increases, $-\beta T_{ij}(t)/\tau_i(t)$ increases, and $dx_i(t)/dt$ increases. This indicates a positive correlation between $\tau_i(t)$ and node failures, that is, the larger the self-recovery factor $\tau_i(t)$, the less likely the node is to fail. When the node is unstable, it can gradually recover to a stable state through the self-recovery mechanism. The self-recovery capability is related to the stability of station facilities and equipment, optimization of passenger flow organization mode, and reasonable allocation of emergency rescue resources.

(3) Mechanism of fault propagation between nodes.

Equation (19) shows the fault propagation mechanism, representing the trend and capability of the fault spreading from one node to other nodes.

Buzna [43] pointed out that $\Theta_i[x_i(t)]$ is nonlinear and is a smooth monotonic increasing function that is very similar to the sigmoid function commonly used in neural networks. Therefore, a sigmoid function is used to represent $\Theta_i[x_i(t)]$, as shown in Eq. (22):

$$\Theta_i\left[x_i(t)\right] = \frac{1-\exp\left[-\alpha x_i(t)\right]}{1+\exp\left\{-\alpha\left[x_i(t)-\theta_i\right]\right\}} \tag{22}$$

Here, α is the gain parameter, and θ_i is the threshold value for the tolerable fault state of the station. When $x_i(t) = 0$, then $\Theta_i[x_i(t)] = 0$, and when $x_i(t) > 0$, the node's fault state value exceeds the station's tolerable fault state threshold, and the unstable effect of the node is transmitted to adjacent stations through trains and lines. The relationship between $x_i(t)$,$\Theta_i[x_i(t)]$ and α is usually that the larger α is, the faster the curve of $\Theta_i[x_i(t)]$ changes, and the greater the sensitivity to the changes;β is the gain parameter.$M_{ij}(t)$ represents the degree of influence and connectivity strength between node S_i and node S_j at time t, determined by the coupling relationship between nodes. This includes factors such as train departure intervals, topological characteristics

of inter-station lines, intensity of information exchange between stations, and so on.

We usually assume that $M_{ij}(t) = 1$ [43]; when $M_{ij}(t)$ increases, $dx_i(t)/dt$ increases, which indicates the positive correlation between $M_{ij}(t)$ and node failures.

$T_{ij}(t)$ is the time delay factor between node S_i and node S_j at time t, which can be considered the transmission time between station S_i and station S_j, and $w_{e_{ij}}(t)$ is measured by the time-weight value of the shortest path on edge e_{ij}. $f[O_i(t)]$ is the degree function of station S_i, and the outdegree value $O_i(t)$ represents the direct impact of station S_i on other adjacent stations at time t. When $O_i(t)$ increases, then $f[O_i(t)]$ increases, $dx_i(t)/dt$ reduces, which indicates a negative correlation $O_i(t)$ between and node failures. $f[O_i(t)]$ can be expressed as Eq. (23), where $a = 1$, $b = 10$ [43]:

$$f\left[O_i(t)\right] = \frac{aO_i(t)}{1 + bO_i(t)} \tag{23}$$

(4) Other parameters

$\zeta_i(t)$ represents random noise disturbance within the station, typically following a uniform or normal distribution. This assumes that it follows a uniform distribution, that is $\zeta_i(t) \sim U(0, \Delta u)$, $\Delta u = 0.001$.

5 Case Study

5.1 Identification and Weighting of Key Hazards

5.1.1 Identification of Key Hazards

The metro operation dispatch log is a text description of the actions, movements, and event status of station personnel when certain situations occur during the subway operation process, which is recorded in real-time by station dispatch personnel. It includes descriptions related to both normal and hazardous events. The log does not require a unified format and covers various aspects of operation data, including power supply, signal, vehicle, passenger transportation, dispatch, and objective environment under various operating conditions. The original data used for text mining is the subway operation dispatch log of a certain subway company's operating line from 2018 to 2020. The control center of the line records the operation log, which includes the station name, date of occurrence, detailed time, event description, vehicle number, vehicle type, reporting time, reporting personnel, detailed repair time, event cause, cause subdivision, vehicle delay time, number of late departures, number of late arrivals, and vehicle operation adjustments. The "content" field, which is a core field for objectively recording event content is included. Due to the limited length of the article and the confidentiality of the data, only selected contents from the original data are included.

The experimental platform is an Intel® Core™ i5-10210U CPU, 2.11 GHz, with 16 GB of RAM, running on a 64-bit Windows 10 operating system. Data processing was carried out in Pycharm software using Python statements.

Mining of hazards based on the improved Apriori algorithm

(1) Data preprocessing

Firstly, the interference data in the original data of the subway operation log were cleaned up, and 38,465 pieces of data related to operational risk events were obtained from 102,834 pieces of original data. The results of removing interference data are shown in Table 6.

The Jieba library is used to segment and deactivate the data after de-interference, and word vector embedding is performed on the obtained data to obtain the preprocessed data as shown in Table 7.

(2) Data analysis

The preprocessed operation log data are input into the AFP-tree algorithm, and the Pycharm operation tool is used for data analysis. The final transaction frequency pattern is obtained. Because the final transaction frequency pattern is large, only some patterns are selected for display, as shown in Table 8.

The confidence of each association rule was calculated by the AFP-tree algorithm. To address this deficiency, a visualization method was introduced using R language and the RStudio

Table 6 Results after data cleaning

Original data	Data frequency	Proportion (%)
Normal scheduling records	33,784	32.85
Routine inspection and maintenance	24,032	23.37
Data related to operational risk events caused	38,465	37.41
Other interfering data	6553	6.37
Total	102,834	100.00

Table 7 Data description after pretreatment

ID	Description of operational events
00001	Century Avenue a1/up a2/driver report a3/down a4/emergency handle a5/pulled down a7/driver a8/cut off a9/after a10/bullet train a11/station a12/late departure a13/4 minutes a14/traffic control command a15/its a16/Tangqiao a17/and a18/Nanpu Bridge a19/up a20/passenger a21/through a22
00002	Yanggao Road a23/turn-back line a24/driver report a3/train a25/master controller key a26/unable to open a27/dispatching order a15/its a16/original a28/exit a29/operation a30/arrangement a31/Yanggao Road a23/standby a32/alternative a33/
00003	Yanggao Road a23/car storage line a34/car a35/TCMS fault a36/requirement a37/car a38/stop a39/handle a40/train dispatching a41/arrange a31/car a35/car storage line a34/standby a42/Yanggao Road a23/standby a32/alternate a33/notify a43/driver a8/station a44
…….	……
38464	Guilin Road a98/downlink a4/train a35/driver report a3/train a25/HMI screen a301/display a79/TCMS invalid a260/broadcast a107/use a99/manual broadcast a108/notice a43/maintenance a50/parking a51
38465	Small South Gate a55/Down a4/Train a35/Driver report a3/Train a25/Automatic a125/Manual broadcast a108/Occurrence of a81/Fault a36/Dispatching order a15/Its a16/Operation a30/Notification a43/Station a44/Broadcast a107/Notification a43/Maintenance a50/Attendance a51

Table 8 Final transaction frequency pattern

Item	Frequent mode
Driver	……
Downstream	(Driver, Downbound: 25014)
Trains	(Driver, Downbound, trains: 21618)
Upstream	(Driver, Upbound: 160458)
Delay	(Driver, Downbound, trains, adjustment, delay: 2391) (Driver, downstream, dispatching, adjustment, delay: 1587)
……	……
ATC fault	(Driver, Upbound, dispatching command, display, train, depot return, ATC fault: 3458)
Door fault	(Driver, Upbound, Trains, Passengers, Re-started, Carriage, Delayed departure, Faulty door: 3762)
(Driver, Downbound, Trains, Carriage, Delayed departure, Passengers, Faulty door: 2973)	
Departure lately	(Driver, Downbound, Trains, adjustment, operation, train dispatching, fault, late departure of train: 2653) (peak, cause, large number of people, late departure of train: 1347)
Clip	(Attendant, platform screen doors, display, stop, check, clip: 2482) (driver, infrared alarm, switch, notice, bullet train, clip: 1961)
Catenary	(Driver, confirmed, impact, high-speed switch, operation, foreign object, overhead contact wire: 1265)
Platform screen doors	(Duty Officer, Unable to Open, Malfunction, Passenger, Train, Handling, Notification, Carriage, Screen Door Shielding: 4673)
False alarm of fire	(Driver, display, three poles, passengers, gate, fall down, enter the station, cause, reset, false alarm of fire alarm: 968)

platform for data analysis. Initially, a low support of 0.02 and a confidence of 0.1 were set, resulting in 12,257 association rules as shown in Fig. 9a. Most of the association rules had relative support ranging from 0.02 to 0.15 and confidence ranging from 0.6 to 0.1. Some association rules were outside this range and were indicated by lighter colors, suggesting that they had insufficient lift and may be ineffective rules or have insufficient association. Based on this analysis, the parameters were readjusted to further filter effective association rules, resulting in 216 association rules as shown in Fig. 9b, with a minimum support of 0.1 and a confidence threshold of 0.8. These 216 association rules were then screened to remove invalid rules with lift less than 1 and rules that did not contain hazardous sources, resulting in 79 valid association rules and 27 key hazardous sources extracted from them.

Filtering the 216 association rules by removing ineffective rules with a lift less than 1 and those that do not contain hazardous sources resulted in 79 effective association rules, from which 27 key hazard sources were identified. The higher the confidence level, the greater the probability that

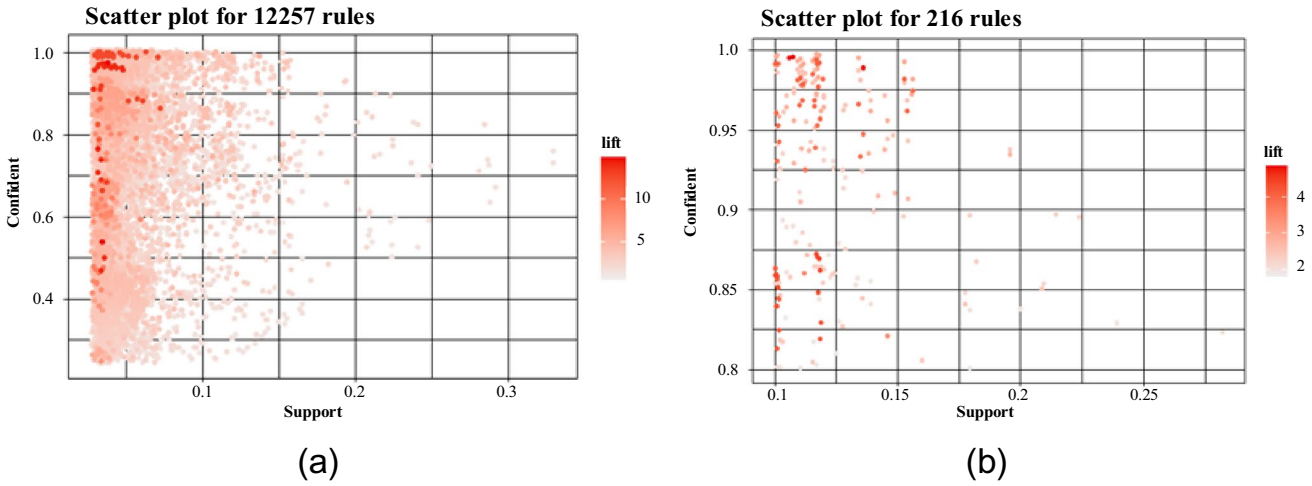

Fig. 9 Comparison of association rules before and after adding sequential constraints

Table 9 Confidence value of key hazard sources

Hazard	Mass passenger flow	Door fault	Screen door clamp	Broadcast failure	VOBC (vehicle on-board computer) crashes	EB (emergency braking)
Confidence value	0.9535	0.9365	0.9322	0.9300	0.9288	0.9274
Hazard	Air- conditioner failure	Wireless communication failure	Communication failure	Screen failure	Mode loss	Wheel diameter loss
Confidence value	0.9207	0.9199	0.9031	0.9003	0.8989	0.8944
Hazard	Location loss	WSP (wheel speed sensor)	Brake failure	Auxiliary inverter	ATS (automatic train supervision) failure	Display
Confidence value	0.8944	0.8932	0.8876	0.8622	0.8599	0.8597
Hazard	TOD (transit-oriented development) failure	Red band	MAU (movement authority unit) lost	Signal failure	Catenary fault	Passengers
Confidence value	0.8501	0.8356	0.8233	0.8233	0.8196	0.8166
Hazard	ATP (automatic train protection) failure	Braking (others)	DT (data transmission) failure			
Confidence value	0.8103	0.8076	0.8041			

the hazardous source will lead to a risk event, which requires focused prevention and control. The confidence levels of each hazardous source are shown in Table 9.

The 27 identified hazardous sources are mainly concentrated in the areas of rolling stock, signaling, and external factors, and should be subjected to focused control measures. Specifically, high-confidence and high-risk facilities and components should be controlled. Meanwhile, attention should be paid to the existence of objective hazards and the dynamics of trains and passengers at stations during operation.

5.1.2 Weighting of Key Hazard Sources

To ensure readability and due to space limitations, this section focuses on the eight key risk sources with the highest confidence levels identified in Sect. 5.1.1 and conducts a weight analysis for these sources as well as subsequent analysis on the propagation of chain failures in later sections. The specific process for assigning weights to key risk sources is described below.

Table 10 Relative importance value of hazard sources

Mass passenger	Door fault	Brake failure	Abnormal passenger behavior	Display failure	Automatic train protection (ATP) failure	Automatic train supervision (ATS) failure	Screen door clamp
0.8582	0.6556	0.5402	0.4083	0.4862	0.7101	0.3440	0.2507

Table 11 Weighting results of key hazard sources

Mass passenger flow	Automatic train protection (ATP) fault	Door fault	Brake fault	Display fault	Passengers	Automatic train supervision (ATS) fault	Screen door clamp
0.2018	0.1670	0.1541	0.1270	0.1143	0.0960	0.0809	0.0589

Table 12 The original matrix

Hazard source	First quarter	Second quarter	Third quarter	Fourth quarter
Mass passenger flow	0.9226	0.9647	0.9588	0.9356
Automatic train protection (ATP) fault	0.9466	0.9341	0.9301	0.9213
Door fault	0.8576	0.9013	0.8822	0.8757
Brake fault	0.8136	0.8272	0.8167	0.8097
Display fault	0.9035	0.87335	0.9206	0.9100
Passengers	0.8193	0.8003	0.8265	0.8113
Automatic train supervision (ATS) fault	0.8509	0.8499	0.8521	0.8323
Screen door clamp	0.9301	0.9045	0.9521	0.9377

Table 13 Standardization of key hazard source scores

Season	x1	x2	x3	x4	x5	x6	x7	x8
1	0.0000	1.0000	0.0000	0.2229	0.6381	0.7252	0.9394	0.5378
2	1.0000	0.5059	1.0000	1.0000	0.0000	0.0000	0.8889	0.0000
3	0.8599	0.3478	0.5629	0.4000	1.0000	1.0000	1.0000	1.0000
4	0.3088	0.0000	0.4142	0.0000	0.7757	0.4198	0.0000	0.6975

(1) Subjective weighting

The subjective weighting section adopts the previously mentioned ordinal relationship method. The first step of the ordinal relationship method generally involves expert ranking of multiple factors' relative importance, followed by assigning relative importance values to adjacent factors. The confidence values of each hazard source can be obtained in the data mining process. Therefore, in the expert scoring process, the confidence values were combined with the estimated cost losses when each hazard source occurred, and the relative importance ranking of each hazard source was ultimately determined and weighted.

The estimated relative cost losses (C) when each hazard source occurs are determined by experts and are set between 0 and 1. C is multiplied by the confidence value to obtain the relative importance value of each hazard source, as shown in Table 10.

The relative importance ranking of the eight key hazards can be determined from Table 10 as follows:

Crowding > ATP failure > Door malfunction > Brake malfunction > Display malfunction > Passenger > ATS failure > Shield door clamp/person

The data from Table 10 are used in Eqs. (5)–(9) and normalized to obtain the results of subjective weighting for the key hazards, as shown in Table 11.

That is, the subjective weight W_l is:

$$W_l = (0.2018, 0.1670, 0.1541, 0.1270, 0.1143, 0.0960, 0.0809, 0.0589)$$

Table 14 Single P value of each hazard source

Season	x1	x2	x3	x4	x5	x6	x7	x8
1	0.0000	0.5394	0.0000	0.1373	0.2644	0.3381	0.3321	0.2406
2	0.4611	0.2729	0.5058	0.6162	0.0000	0.0000	0.3143	0.0000
3	0.3965	0.1876	0.2847	0.2465	0.4143	0.4662	0.3536	0.4474
4	0.1424	0.0000	0.2095	0.0000	0.3214	0.1957	0.0000	0.3120

Table 15 Information entropy of key hazard sources

	x1	x2	x3	x4	x5	x6	x7	x8
Information Entropy	0.7223	0.7223	0.7429	0.6609	0.7802	0.7514	0.7916	0.7690

Table 16 Objective weights of key hazard sources

	w1	w2	w3	w4	w5	w6	w7	w8
weight	0.1349	0.1348	0.1248	0.1647	0.1067	0.1207	0.1012	0.1122

(2) Objective weighting

(1) Build the original matrix

The quarterly original matrix for each hazard source is constructed in Table 12:

(2) Calculate the standardized score table for eight key hazard sources in four quarters, as shown in Table 13
(3) Calculate the individual P value of each hazard source from Eq. (13), as shown in Table 14. Further, obtain the information entropy of each key hazard source from Eq. (12), as shown in Table 15:
(4) Finally, calculate the objective weight of the required key hazard sources according to Eq. (14), as shown in Table 16:

The objective weight W_l:

$$w_l'' = (0.1349, 0.1348, 0.1248, 0.1647, 0.1067, 0.1207, 0.1012, 0.1122)$$

(3) Combinatorial weighting

Combining subjective and objective weighting methods and substituting them into Eq. (15), the combined weight is obtained as follows:

$$w_l = \frac{W_l' \times W_l''}{\sum_{l=1}^{n} W_l' \times W_l''} = (0.21, 0.18, 0.15, 0.16, 0.09, 0.09, 0.06, 0.05)$$

Fig. 10 ArcMap mapping of the Shanghai Metro network

5.2 Transmission of Subway Network Cascading Failures

5.2.1 Analysis of Complex Topological Network Characteristics of the Shanghai Metro

The network propagation model of key hazard sources constructed in Sect. 4 was applied to analyze the complex network characteristics of the Shanghai subway and to build a complex network model of the subway. A network crawler was used to extract the line and station coordinate data of the Shanghai subway network from Baidu Maps, and the results were saved as .shp files. ArcMap software in ArcGIS was used to visualize and display the data, as shown in Fig. 10.

Based on the key hazard identification method in Sec. 4, a network fault propagation model was constructed, and the complex network characteristics of the Shanghai Metro were analyzed to build a complex network model. A network crawler was used to crawl the line and station coordinate data of the Shanghai Metro network from Baidu Maps, and the

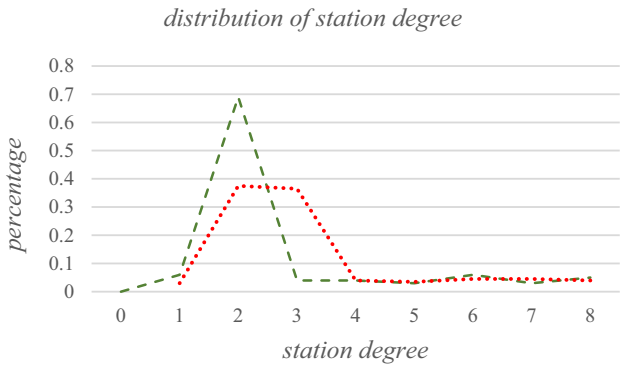

Fig. 11 Probability statistics of Shanghai Metro station degrees

results were saved as a .shp file. ArcMap software in ArcGis was used to visualize the results, as shown in Fig. 10.

Using the spatial join function in ArcMap, the connection relationships in the geographic coordinates were transformed into an adjacency matrix and stored in tabular form, resulting in a 508×508 adjacency matrix of the Shanghai Metro network. Under the Space-L method, the degree of a subway station was used to represent the number of adjacent stations around that station. For example, at Xujiahui Station, you can transfer to Line 1 to reach Shanghai Stadium and Hengshan Road Station, transfer to Line 11 to reach Jiaotong University and Shanghai Swimming Center Station, and transfer to Line 9 to reach Yishan Road and Zhaojiabang Road Station. There are six adjacent subway stations around Xujiahui Station. Therefore, the degree value of Xujiahui Station is 6. The degree values of 508 stations were calculated using the station adjacency matrix. Tables 5, 6, 7, 8, 9, 10, 11 and 12 summarize the degree values of some representative stations, most of which have high transfer convenience (such as the large transfer hub People's Square Station and Century Avenue Station). The average degree of the entire Shanghai Metro network was calculated to be 2.38, indicating that there are about two to three stations that can be directly reached around each station, and the convenience of traveling by subway is relatively high. To better quantify the probability distribution of degree values of Shanghai Metro stations, the distribution of station degrees was drawn using Python software, as shown in Fig. 11. The probability of stations with a degree value of 2 is as high as 71% (Table 17).

Using Python, the node degree distribution of the subway network was calculated and the cumulative degree distribution graph was plotted, as shown in Fig. 12. The horizontal axis represents the logarithm of the node degree value, while the vertical axis represents the logarithm of the probability of having a degree value greater than the corresponding node degree value. From Fig. 12, it can be seen that the sample stations exhibit a slow decay in double logarithmic coordinates, showing a small-scale, scale-free regime. The cumulative degree distribution graph of the subway stations in Shanghai basically conforms to the power law characteristic. Therefore, in L space, the degree distribution of the Shanghai subway network can be roughly described by a power law distribution, indicating that the Shanghai subway network belongs to a scale-free network.

5.2.2 *Analysis of Chain Failure Propagation in the Shanghai Metro Network*

The OD matrix was constructed using the automatic ticketing system of the Shanghai Metro on a certain day. The train timetable, maximum passenger capacity, transfer distance, average walking speed of passengers, train departure interval, maximum queuing capacity, number of escalators and stairs, and escalator and stair output rates for each station

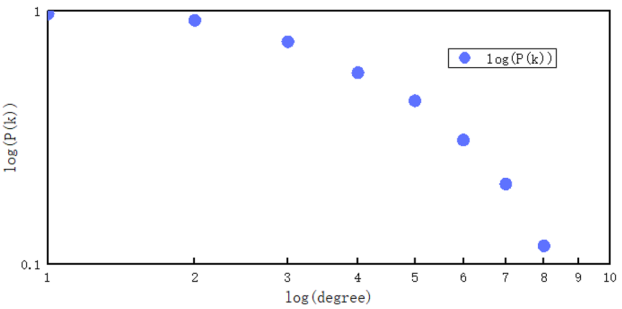

Fig. 12 Cumulative degree distribution in double logarithmic coordinates

Table 17 Example of partial station degrees

Station no.	Station	Degree	Station no.	Station	Degree
1	Century Avenue	8	8	Jing'an Temple	6
2	Longyang Road	8	9	Caoyang Road	6
3	Xujiahui	6	10	Oriental Sports Center	5
4	People's Square	6	11	Changshu Road	5
5	Nanjing West Road	6	12	Shanghai South Railway Station	5
6	Hanzhong Road	6	13	Yishan Road	5
7	Shanxi South Road	6	14	Shanghai Railway Station	4

are based on statistics from the Shanghai Metro operating company. Based on the basic data, a weighted network of the Shanghai Metro was constructed. Then, based on the theory of disaster propagation, the size of the failed nodes caused by cascading failures was simulated. The influence of the initial node attribute value $x_i(t)$ on disaster propagation was analyzed, with emphasis on the weighting of critical hazards. Other parameters were set as follows: Δt=2 min, and the total simulation time was 60 min, $M_{ij}(t) = 1$, $\theta_i = 0.9$, $\lambda = 0.15$, $\delta = 5$, a=1, b=10, α=10和 β=0.01 [43].

The basic assumptions are as follows:

(1) Two node failure modes are set, including fixed node failure (representing a pre-defined station as a failure node) and random node failure.
(2) Two types of failure states are set: failure caused by attacks, in which case the initial node attribute value is , meaning the node is completely failed; failure caused by critical hazards, in which case the initial node attribute value is determined by the results of Sect. 5.1.2, representing node failure.
(3) Station failures are divided into two categories: general station failures and transfer station failures.
(4) The self-recovery coefficient has different values set, and.

The specific simulation calculation unit steps are as follows:

Step 1: Initialization. Determine the initial state of the network, network topology, initial OD matrix, travel path, travel time of each travel path, and passenger volume and capacity of each station. Set t=0 to determine the cycle interval time Δ t_0.
Step 2: Update the network topology. Determine the station failures and station types (regular, transfer, or terminal) and, based on Figs. 10 and 11, determine the deletion or retention of nodes and edges in the network, and then update the entire network based on the results.
Step 3: Update the degree function of the stations based on the updated network topology.
Step 4: Update the travel path data. Based on Eqs. (17) and (18), calculate the time cost of each path.
Step 5: Compute the shortest path. Calculate the shortest path based on the calculation steps and Dijkstra's algorithm in Figs. 7 and 8.
Step 6: Update passenger flow at each station. In each cycle interval , the passenger traffic volume of each station can be calculated based on OD data and the shortest path in the network within Δt period.

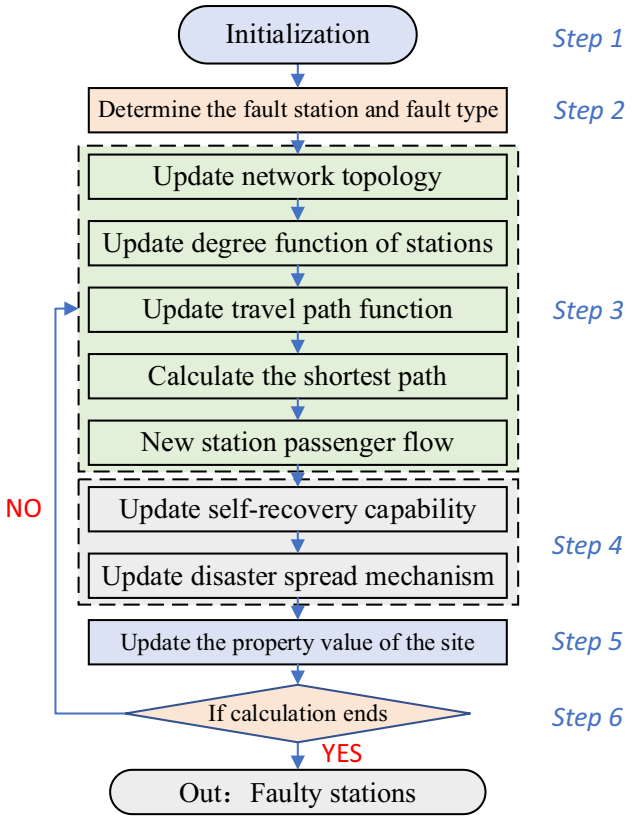

Fig. 13 Flow chart of disaster propagation simulation

Step 7: Update the self-recovery capability of each station, calculate the of the node based on the attribute value Eqs. (20) and (21) and the self-recovery factor .
Step 8: Update the fault propagation mechanism. Calculate the propagation mechanism of disasters in the network based on Eqs. (22) and (23).
Step 9: Update the attribute values of the station. After each cycle interval, update the attribute values according to Eq. (23) and count the number of malfunctioning stations. After each cycle interval Δt, the evolution dynamics of the attribute values over time should satisfy Eq. (19) under the combined action of self-recovery mechanism and fault propagation mechanism.
Step 10: Determine whether the computation is finished. The criteria for judgment are as follows: (1) The set simulation time has ended, and the loop ends. (2) The set simulation time has not ended, but more than half of the nodes in the network have malfunctioned, causing the network to be unable to operate normally, and the cycle ends.

The specific calculation process is shown in Fig. 13:
Therefore, by obtaining and comparing simulation results under different combinations of conditions, and due to the large number of iterations, there are many result graphs, and the length is too long. Only the effect graphs are shown, as

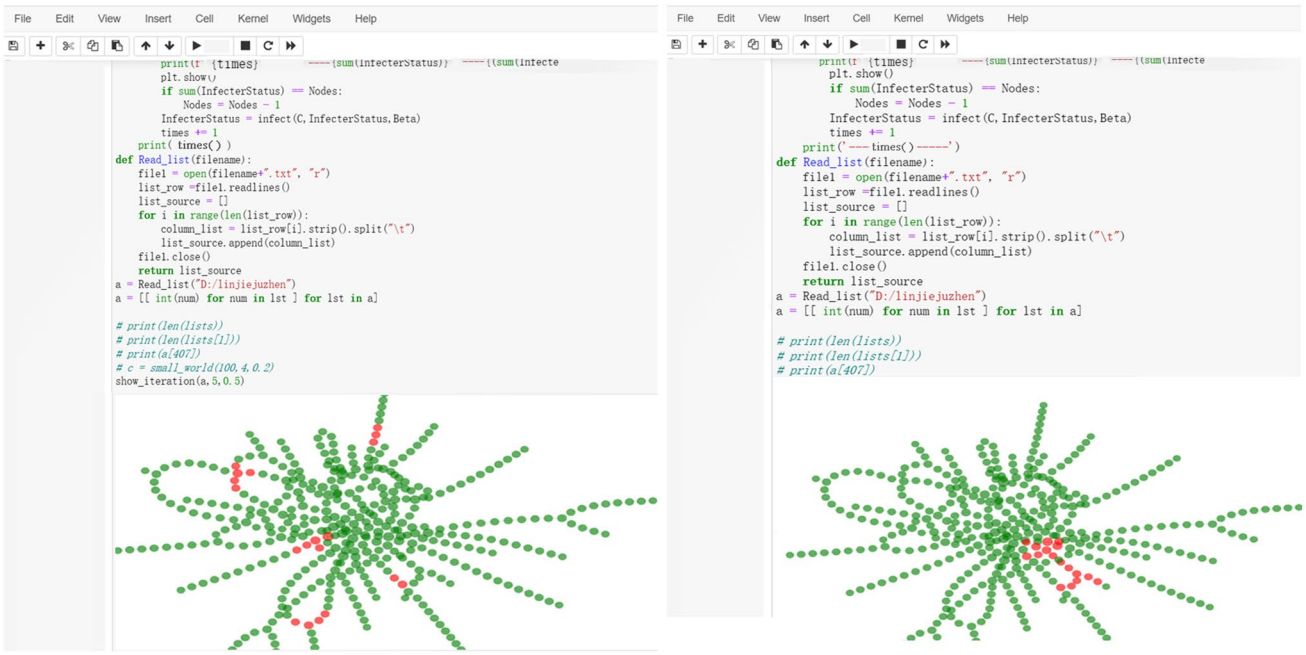

Fig. 14 Fault node simulation process

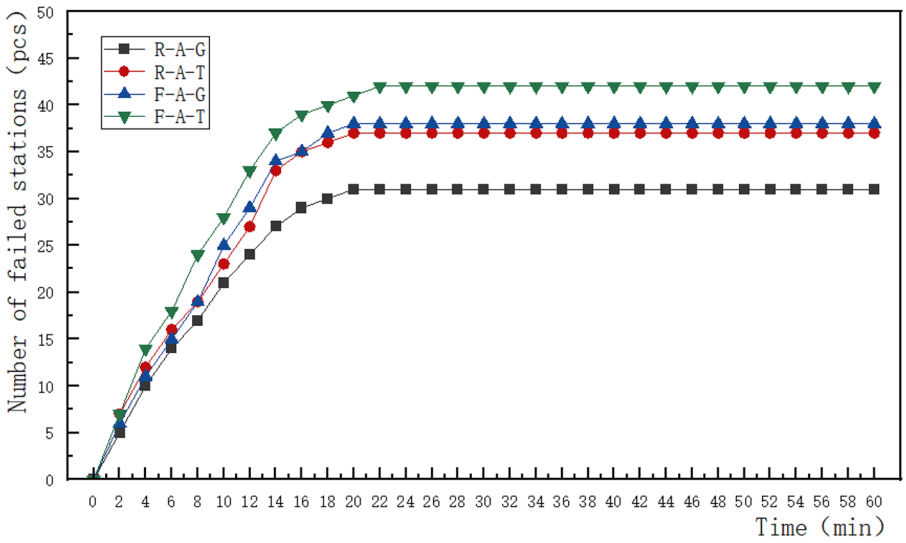

Fig. 15 Scale of failure stations under different simulation scenarios (a)

shown in Fig. 14. The left side shows a screenshot of the simulation process with random faulty nodes, and the right side shows a screenshot of the simulation process with specified faulty nodes.

For convenient comparison and observation, all final simulation results are summarized and presented in the form of line graphs, as shown in Figs. 15, 16, 17, and 18, in which R is random, F is fixed, A is attack, G is general station, T is transfer station, and H is hazard source.

In Fig. 15, for fixed node failure, the initial faulty node is preset as People's Square Station in all combinations. People's Square Station is an interchange station for Shanghai Metro Lines 1, 2, and 8, and plays an important role in the transportation network of Shanghai Metro. After analyzing Fig. 15, we can draw the following conclusions:

(1) With the increase of simulation time, the scale of chain failures of stations under different combinations continues to increase, and remains stable after reaching a certain fixed range. Within 20 min, the Shanghai Metro will lose its normal operational function.

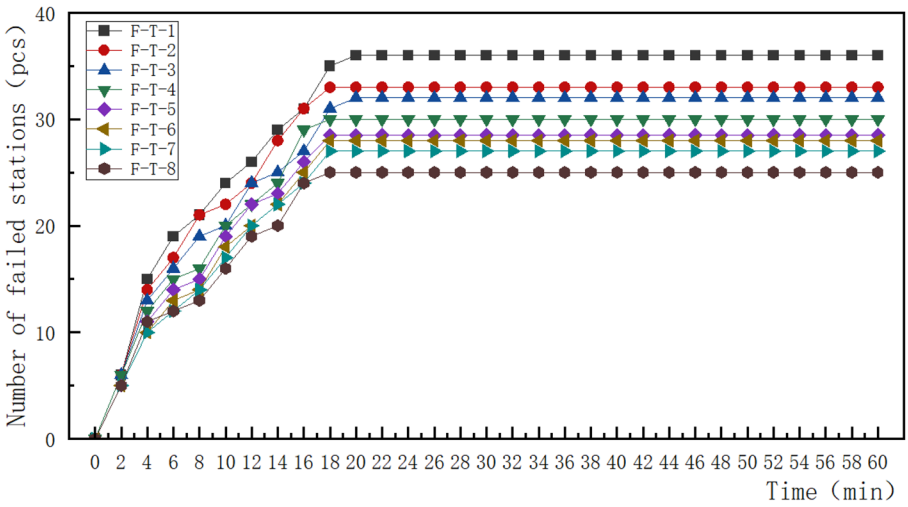

Fig. 16 Scale of failure stations under different simulation scenarios (b)

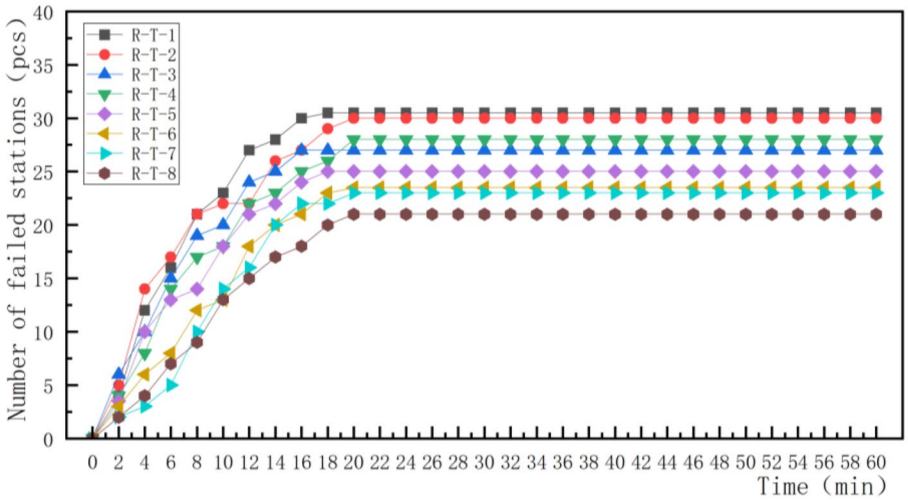

Fig. 17 Scale of failure stations under different simulation scenarios (c)

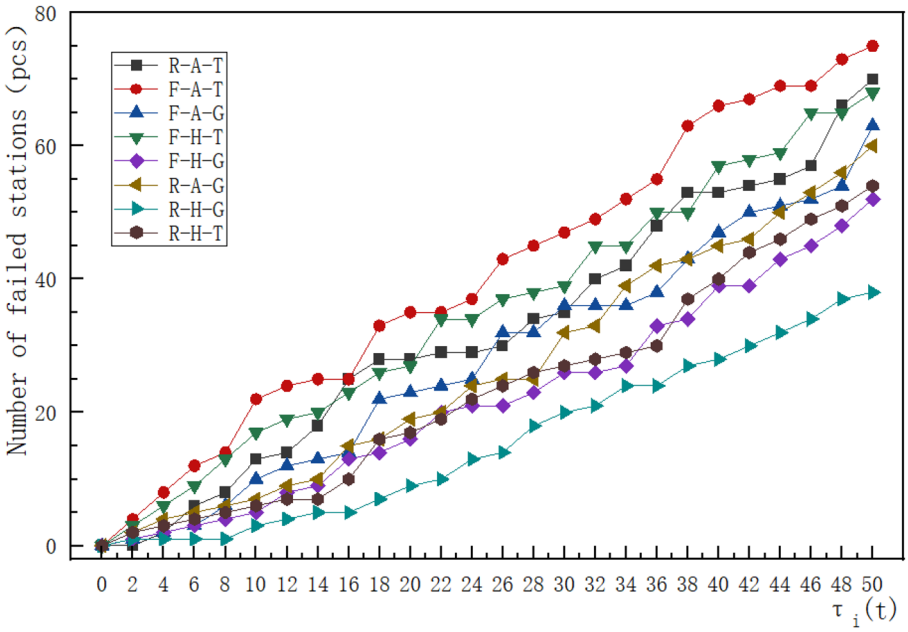

Fig. 18 Scale of failure stations under different simulation scenarios (d)

(2) The number of failed stations under fixed station failure is higher than that under random station failure, indicating that the network topology has stronger robustness under random failure. However, when one or more nodes in the network are intentionally failed, these nodes are easy to fail, and the entire network may be affected due to disaster propagation mechanisms.
(3) The number of failed stations under initial interchange station failure is higher than that under initial ordinary station failure, the cascading failure speed is faster, and the failure propagation range is wider. This means that in the subway transportation network, interchange stations have a substantial influence on the scope and intensity of failure propagation. In terms of transportation operation organization, it is necessary to pay special attention to the land use, bus connections, passenger flow organization, and other aspects of large-scale interchange stations in the network.
(4) The largest scale of failed stations is under fixed attack on interchange stations (with 43 failed stations), which means that the initial failure caused by a fixed interchange station being attacked by terrorism has the greatest impact on the normal operation of the Shanghai Metro network. The smallest scale of failed stations is under random attack on ordinary stations (with 31 stations), which means that the initial failure caused by a random ordinary station being intentionally attacked has the smallest impact on the normal operation of the Shanghai Metro network.

Figure 16 shows the fault scale of Shanghai subway stations with fixed transfer stations under different initial station attribute values for different hazards, while Fig. 17 shows the fault scale of Shanghai subway stations with random transfer stations under different initial station attribute values for different hazards. Observing Figs. 16 and 17, the following conclusions can be drawn:

(1) In Fig. 16, the higher the weight value of the critical hazard, the larger the fault scale and the wider the impact. This indicates that the higher the initial station attribute value, the easier the fault caused by critical hazards is spread in the Shanghai subway network, typically reaching the maximum impact within 20 min. Among them, critical hazard 1 has the widest impact range, involving 37 stations, while critical hazard 8 has the smallest impact range but still affects 25 operating stations.
(2) Figure 17 is similar to Fig. 16, but the overall fault scale is smaller than that in Fig. 16. This indicates that under the same hazard causing station faults, the random occurrence of transfer station faults is generally smaller than the specified station fault scale, which also indicates that the Shanghai subway network has stronger robustness under random settings.
(3) Comparing Figs. 16 and 17, it can be found that critical hazards with higher weight values usually have a more extensive impact on the Shanghai subway network during the simulation process, and usually spread rapidly during two time periods of 2–4 min and 14–18 min, with a significant increase in the fault scale growth rate. Therefore, it is particularly important to be able to handle relevant problems timely and effectively when subway operation safety incidents occur to prevent the escalation of the situation.

In Fig. 18, the influence of different self-recovery factors on the impact of Shanghai Metro transfer stations in the event of attacks or failures caused by critical hazards is fully considered. The following conclusions can be drawn from the figure:

(1) Under the same conditions of initial fault points, as the self-recovery factor $1/\tau_i(t)$ decreases and the self-recovery factor $\tau_i(t)$ increases, the node fault size increases. When the self-recovery ability of a node weakens, the time for the node to recover to a normal state increase. Therefore, during the same period, the larger the value of the node's self-recovery factor, the more difficult it is for the node's unstable state to repair itself, and the disaster propagation mechanism gradually takes the dominant position, resulting in larger node fault sizes.
(2) There is no obvious functional relationship between the fault size and the self-recovery coefficient. When the self-recovery factor is within a certain range, the number of fault stations increases rapidly. For example, for fixed transfer stations under attack, when $\tau_i(t) \in [34, 40]$ and $\tau_i(t) \in [14,18]$, the fault size of multiple simulation combinations increases rapidly.
(3) The self-recovery factor is positively correlated with the cascade failure size. Therefore, when a network experience cascading failure, emergency resources can be configured, the stability of station facilities and equipment can be strengthened, the optimization of passenger flow organization can be carried out, the self-recovery ability can be enhanced, and the self-recovery factor of nodes can be reduced to control cascading failures and reduce the fault size in the network.

6 Conclusions and Further Studies

Based on the association rule theory and text mining methods, this paper proposed the AFP-tree algorithm to mine the operation log data of the Shanghai Metro, identified the key hazards that cause operational risks, conducted a weighted analysis of the key hazards, and determined the proportion of each specific hazard in the subway network operation process. Furthermore, disaster propagation theory was introduced to investigate the propagation time and impact range of each key hazard in the subway network fault propagation process by constructing a subway network chain fault propagation model with the key hazards as the changing indicators. The specific research conclusions are as follows:

(1) By using the AFP-tree algorithm to analyze the operation text data of the Shanghai Metro, 27 key hazards in the Shanghai Metro operation process, including high passenger flow, door malfunction, and shielding door clips, were identified. The importance of the 27 key hazards was sorted according to the confidence level, and the algorithm was proven to be effective in identifying transportation hazards and has practical guiding significance for enterprise operation safety.
(2) Among the key hazards, the eight hazards with the highest confidence level were selected, and the subjective and objective weights were calculated by the sequence relationship method and entropy method, respectively. The eight key hazards, including high passenger flow, automatic train protection (ATP) malfunction, brake malfunction, door malfunction, display malfunction, passenger, automatic train supervision (ATS) malfunction, and shielding door clips, were weighted through combination weighting, laying a foundation for exploring the impact of key hazards on the entire Shanghai Metro network.
(3) A subway network chain fault propagation model was constructed, and the impact of eight key hazard fault propagation of the Shanghai Metro was analyzed in detail. The results showed that when the fixed transfer station was attacked, the fault scale caused by high passenger flow was the largest, with 36 affected stations. When the random transfer station was attacked, the hazard events caused by shielding door clips affected the smallest number of stations, with 21 affected stations. The number of fault stations under different conditions reached the maximum value of 16–20 min, and the specific hazards had different impacts on the subway network. Through example analysis, it was found that under different self-recovery factors, the number of fault stations showed a significant increasing trend when the self-recovery factor was 14–18, indicating a positive correlation between the fault scale and the self-recovery factor.

While the case study establishes the applicability and validity of the methodology presented in this paper, and yields research conclusions with practical value, certain limitations exist in terms of both the number and spatial scope of the collected data cases. Future research endeavors should delve into more comprehensive studies on safety risk factors affecting subway operation, considering the unique circumstances of each city to align with the actual requirements of operational management. Additionally, the exploration of hazard mining within the subway network focuses on evaluating the likelihood of concurrent hazard occurrences and associated risk events. Subsequent efforts will include integrating a dimension analysis of risk loss to further enhance the precision of weight allocation for hazards in the analysis of failure propagation.

Funding This work was supported by the Shanghai Philosophy and Social Science Planning Project under Grant 2022BGL001. We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, and there is no professional or other personal interest of any nature or kind in any product, service or company that could be construed as influencing the position presented in or the review of this manuscript.

References

1. Réka A (2000) Error and attack tolerance of complex networks. Nature 406:378–382
2. Ahmadian N, Lim GJ, Cho J, Bora S (2020) A quantitative approach for assessment and improvement of network resilience. Reliabil Eng Syst Saf 200:106977
3. Esfeh MA, Kattan L (2020) Lam WHK (2020) Compound generalized extreme value distribution for modeling the effects of monthly and seasonal variation on the extreme travel delays for vulnerability analysis of road network. Transp Res Part C Emerg Technol 120:102808
4. Li T, Rong LL (2020) A comprehensive method for the robustness assessment of high speed rail network with operation data: a case in China. Transp Res Part A Policy Pract 132:666–681
5. Ma MH, Wang XH, Xu XH et al (2017) Study on accident hidden danger early warning method based on data mining technology. J Saf Sci Technol 13(07):11–17

6. Niu Y, Fan YX, Gao Y (2019) Topic extraction of chemical production accidents based on data mining. J Saf Sci Technol 15(10):165–170
7. Li CD, Li WB, Cao CJ et al (2017) Analysis of urban public safety and its risk hotspots based on news search. J Saf Sci Technol 13(08):73–79
8. Luo WH, Cai FT, Wu CN et al (2021) Identification model of road transport safety risk sources based on text mining. J Southwest Jiaotong Univ 56(01):147–152
9. Li J, Wang YF (2020) Analysis of causes network of high altitude fall accidents in construction based on text mining. J Saf Environ 20(04):1284–1290
10. Xue NN, Zhang JR, Zhang W et al (2021) Study on unsafe behavior of construction workers and its influencing factors based on text mining. Saf Environ Eng 28(02):59–65
11. Wu J, Jiang FC, Yao HJ et al (2018) Analysis of the causes and risk prediction of inland ship collision accidents based on text mining. Transp Inf Saf 36(03):8–18
12. Fa ZW, Li XC, Qiu ZX (2021) From correlation to causality: Path analysis of accident-causing factors in coal mines from the perspective of human, machinery, environment and management. Resour Policy 73:102157
13. Na X, Ling M, Liu Q, Li W, Deng Y (2021) An improved text mining approach to extract safety risk factors from construction accident reports. Saf Sci 138:105216
14. Chu C-Y, Park K, Kremer GE (2020) A global supply chain risk management framework: an application of text-mining to identify region-specific supply chain risks. Adv Eng Inform 45:101053
15. Zhao LT, Guo SQ, Wang Y (2019) Oil market risk factor identification based on text mining technology. Energy Procedia 158:3589–3595
16. Qiu Z, Liu Q, Li X, Zhang J, Zhang Y (2021) Construction and analysis of a coal mine accident causation network based on text mining. Process Saf Environ Protect 153:320–328
17. Ahadh A, Binish GV, Srinivasan R (2021) Text mining of accident reports using semi-supervised keyword extraction and topic modeling. Process Saf Environ Protect 155:455–465
18. Liu X, Saat MR, Barkan CPL (2012) Analysis of causes of major train derailment and their effect on accident rates. Transp Res Rec J Transp Res Board 2289(1):154–163
19. Wang H, Tian Y, Yin H (2021) Correlation analysis of external environment risk factors for high-speed railway derailment based on unstructured data. J Adv Transp 2021:1–11
20. Wang XF, Li X, Chen GR (2006) Complex network theory and its applications. Tsinghua University Press
21. Li Q (2017) Evaluation of node importance and cascading failure resilience of urban rail transit networks. Beijing Jiaotong University
22. Watts DJ (2002) A simple model of global cascades on random networks. Proc Natl Acad Sci USA 99(9):66–71
23. Bak P, Tang C, Wiesenfeld K (1990) Self-organized criticality. Physica A 163(1):403–409
24. Linton CF (1978) Centrality in social networks conceptual clarification. Soc Netw 1(3):215–239
25. Wang JW, Rong LL (2008) A model for cascading failures in scale-free networks with a breakdown probability. Physica A Stat Mech its Appl 388(7):1289–1298
26. Li P, Wang BH, Sun H (2008) A limited resource model of fault-tolerant capability against cascading failure of complex network. Eur Phys J B Condens Matter 62(1):101–104
27. Duan DL, Wu J, Deng HZ (2013) A cascading failure model for complex networks based on adjustable load redistribution. Syst Eng Theory Practice 33(1):203–208
28. Fang X, Yang Q, Yan W (2014) Modeling and analysis of cascading failure in directed complex networks. Saf Sci 65(3):1–9
29. Ma D (2017) Research on the mechanism of cascading failures in rail transit networks based on complex networks. Lanzhou Jiaotong University
30. Zhihao J, Jinlong M, Jianjun W (2018) Cascading failure model for improving the robustness of scale-free networks. Int J Mod Phys C 29(06):185–196
31. Li C, Zhang S, Yang Z et al (2019) Simulation of cascading resilience of urban passenger transport network under intentional attack. J Transp Syst Eng Inf Technol 19(02):14–21
32. Buzna L, Peters K, Helbing D (2006) Modelling the dynamics of disaster spreading in networks. Physica A 363(1):132–140
33. Buzna L, Peters K, Ammoser H (2006) Efficient response to cascading disaster spreading. Phys Rev E Stat Nonlinear Soft Matter Phys 75(5–2):056107
34. Hu ZH, Sheng ZH (2015) Disaster spread simulation and rescue time optimization in a resource network. Inf Sci 298:118–135
35. Yi T, Zhu QX (2014) Simulation and application of a disaster spread model in Chemical Disaster Network - ScienceDirect. J Loss Prev Process Ind 27(1):130–137
36. Ouyang M, Fei Q, Yu M (2008) Evaluation and improvement of disaster spreading model based on complex networks. Acta Physica Sinica 57(11):6763–6770
37. Ouyang M, Fei Q, Yu MH (2009) Effects of redundant systems on controlling the disaster spreading in networks. Simul Modell Practice Theory 17:390–397
38. Weng WG, Ni SJ, Shen SF et al (2007) Study on the dynamics of disaster spread on complex networks. J Phys 04:1938–1943
39. Xiao WJ, Zhang Q (2018) Modeling and simulation of congestion propagation based on disaster propagation theory. J Railway Sci Eng 15(06):1593–1600
40. Han JW (2007) Data mining: concepts and techniques (3rd Edition). Machinery Industry Press, Beijing
41. Zheng JG (2019) Research and application of data mining. Yunnan University Press, Yunnan
42. Wen F, Huang HL, Li TD et al (2020) Rapid mining algorithm for library data based on FP-growth association rules. J Chongqing Univ Technol 34(06):189–194
43. Li NN, He ZY (2009) Comprehensive evaluation of power quality based on combination of subjective and objective weights. Power Syst Technol 33(06):55–61
44. Yang Y (2006) Analysis of weighting methods in multi-index comprehensive evaluation. Stat Decis 13:17–19