Research on Aircraft Runway Accident Investigation Report Based on LDA-Apriori Algorithm

Jianping Bao

821099895@qq.com

Civil Aviation Safety Engineering College, Civil Aviation Flight University of China,Guanghan Sichuan 618307,China

Abstract. In order to effectively prevent the occurrence of aircraft runway accidents, the improved text topic model (LDA) and association algorithm (Apriori) were used to dig out the causes of aircraft runway safety accidents and analyze the correlation between them. Firstly, 181 accident investigation reports were obtained after collecting nearly 20 years of aircraft runway accident investigation reports. Secondly, the collected text data was cleaned and the text was processed by the TF-IDF algorithm in the improved LDA model, and the ten topics of the accident text, the key feature words representing each topic and the probability distribution in the text were obtained. Finally, the association rules of each topic and subject term are analyzed, and the strong association rules are excavated to understand the key causal factors of aircraft runway accidents. The results show that the correlation path of the five nodes, including pilot status, pilot operation behavior, system failure, landing gear, and management system, is complex and highly correlated with other nodes, which is the key causal factor of aircraft runway accidents, and landing gear collapse is the most common type of accident among aircraft runway accidents, and the causative chain "pilot state→operation error→related system failure→landing gear collapse→accident" extracted from this paper.

Keywords: aircraft; accident investigation report; topic models (LDA); association algorithm (apriori); cause of the accident

1. Introduction

Airport runway safety is related to the safety of aircraft take-off and landing. With the increase of passenger traffic, the number of airports with multiple runways is increasing, and a series of runway safety issues related to aircraft operation have gradually become a hot issue in the aviation industry and academia. The take-off and approach landing stages are the most risky periods in the whole flight process[1], The pilot's task intensity is the heaviest in the take-off and approach landing stages, and the pilot's efficiency curve shows a downward trend during the whole flight process. When the task intensity curve intersects with the pilot's efficiency curve, runway safety problems will occur.

In order to effectively prevent airstrip accidents and improve safety management level, relevant scholars conducted relevant research and analysis on previous runway accidents from the perspectives of accident causes and association rules analysis, so as to find potential risk factors as much as possible and provide corresponding countermeasures. Based on the analysis model of aircraft landing off-runway event integrating fault tree and Bayesian network, Huo Zhiqin et

PMBDA 2023, December 15-17, Nanjing, People's Republic of China

DOI 10.4108/eai.15-12-2023.2345370

al,determined the main reasons for aircraft landing off-runway [2]. Lu Qinqian and Yu Tingting[3] took the multiple entry runway intrusion incidents at Beijing Capital International Airport as a case study to analyze and put forward improvement measures and suggestions for this type of runway intrusion to reduce the risk of runway intrusion. Zhao Xianli[4] determined the life cycle of runway safety risk evolution based on evolution curve and proposed a life cycle correction model for runway safety risk evolution. By combining system dynamics and game theory, an evolutionary game model of airport runway safety stakeholders is constructed. Liu Mengna[5] used deep learning to analyze aviation safety accident reports. Experiments proved that the neural network classification model based on deep learning has low dependence on artificial feature selection and can realize automatic extraction of text features. Its overall classification effect is better than traditional machine learning, which optimizes the classification effect. Zhou Xiuting[6] used the project-based management theory to treat a single unsafe incident as a project and batch unsafe incident as a project set, and then combined the project risk management theory and method with text mining technology to apply the unsafe incident project set, complete the risk identification, analysis, response and monitoring, and build a complete civil aviation safety risk management and control system. Ma Ting[7] collected 90 cases of accident reports and obtained 8 major risk factors and 17 general risk factors through text mining and K-means clustering, which provided valuable reference information for future flight safety from three aspects of human, environment and equipment.

In the field of text mining, Blei proposes[8]Latent Dirichlet Allocation (LDA) model, which can extract text theme words, in order to more effectively mine key information of text and extract text theme words based on the need to extract theme words. And calculate the degree of association between the subject and the subject word, and use the numerical value to represent the subject, which is more direct and objective. Latent Dirichlet Allocation (LDA) is a document topic generation model, also known as a three-tier Bayesian probability model, which contains a three-tier structure of words, topics, and documents. The so-called generation model, that is to say, we think that each word of an article is through the process of "selecting a certain topic with a certain probability, and selecting a certain word from this topic with a certain probability", and the document to the topic follows a polynomial distribution, and the topic to the word follows a polynomial distribution. Relevant scholars use LDA model to extract the themes and subject words of the text, and identify the extracted topics so as to obtain the topic distribution probability and the core of the text. Griffiths[9]et al. built an LDA model to analyze abstracts of scientific and technological papers published in the Proceedings of the National Academy of Sciences over the past 20 years, and mined the topics and classified the hot topics and non-hot topics through the LDA model. Huang Jingwei[10]et al. used LDA model to extract users' operational behaviors from very large security monitoring data and form a set, so as to discover users' possible hidden intentions and thus improve the organization's network security protection mechanism. Huang Guimin[11]et al. improved the LDA model and proposed an improved SLDA model on the basis of the original model. The SLDA topic model added a sentence hierarchy analysis process, which made up for the shortcomings of the bag-based method that ignored inter-word relations and overcame the shortcomings of the model that required a large number of training texts. Caicai Zhang[12]et al. proposed to combine deep learning with traditional machine learning methods, that is, the combination of PCA and LDA convolutional networks, to form a PLDANet model, and verified the practicality of the model through experiments.

In China, relevant scholars have also conducted a lot of research on the LDA model, and according to their own research needs, made corresponding improvements on the basis of the original model, and built a more applicable LDA model. Sun Jiazhi [13] collected 44 major coal mine accident investigation reports published by the coal mine safety supervision bureaus of various provinces, mined and analyzed them by machine learning method, obtained their theme distribution probability, and then obtained important information such as the main theme and key hidden dangers of the accident report, thus providing corresponding management measures for coal mine safety management. Reduce the probability of coal mine safety accidents. Yang Lichao [14] et al,took the data of water traffic accidents in China's coastal waters from 2015 to 2020 as samples, analyzed the characteristics of water traffic accidents in China's coastal waters, and dug out the potential mapping relationship between accident attributes and causative attributes. By combining text mining with complex network theory, Zheng Binbin et al, [15] identified the key causes of gas accidents and the set of causes of accidents, and found that other leaks are common types of gas accidents, which are mainly strongly related to pipeline breakage and open valve. Yan Ling [16] helped LDA-Cor Ex mixed theme model to mine traffic accident reports, found the cause-causing themes and analyzed the evolution characteristics of the themes, so as to identify the key cause-causing chains of road traffic accidents, and put forward relevant suggestions from three aspects: source chain breaking, cause-causing prevention and crisis learning in combination with the chain breaking disaster reduction theory.

Text mining technology has been widely used in books, emotional problems, satisfaction surveys, etc. Relevant scholars have also conducted analysis and research on airstrip safety accidents through LDA model. However, due to the large amount of text data, many themes and subject words have been extracted, which will cause interference in the analysis of themes and subject words. The accuracy of analysis results is affected. In view of this, the author proposes to use the combination of text mining technology and association rules, and use the Apriori algorithm to calculate the association rules for the topics and subject words extracted from the LDA model, and screen the accident factors for obtaining strong association rules. The main causes of airstrip accidents are found through the analysis of the accident causes of the strong association rule path, which provides a theoretical basis for the prevention and safety management of airstrip accidents.

2. Text mining of runway accident investigation report

2.1 Text word frequency extraction

The Aviation Herald and the international Aviation Safety Network database collected information on airstrip accidents that occurred around the world in the past ten years, and obtained 181 investigation reports of runway accidents after screening and deleting relevant duplicate and inconsistent data.

First of all, the text is segmtioned by NLTK library, before which it is necessary to add a personal dictionary and extend a deactivated dictionary. Since a large number of aviation professional words have not been entered into the default dictionary of the system, in order to prevent the reliability of the results from being affected due to the failure to identify them during word segmentation, it is necessary to manually sort out the relevant professional words. Therefore, the relevant professional words were sorted out in the early stage, as shown in Table 1 below.

Table 1. Professional word

Number	Professional word
1	RVR
2	CAT
3	Wind-sock
4	Low-level windshare
…	…
2850	Build-up
2851	Aviation-grade petrol

On the other hand, the extended stop words are added, and the words irrelevant to the study are screened out, such as airport name, place name, country name, etc. The extended stop words are sorted out as shown in Table 2 below.

Table 2. Stop word

Number	Stop word
1	Airport
2	Deadhorse
3	Badami
4	Departure
5	Contributory
…	…
1058	Hr-ayy

Secondly, after the personal dictionary and extended deactivated dictionary are imported into the NLTK database, word segmentation is performed on the corpus text. Word frequency statistics are performed on the word segmentation results on the one hand, and word cloud map is displayed on the other hand. The word cloud map results are shown in Figure 1 below

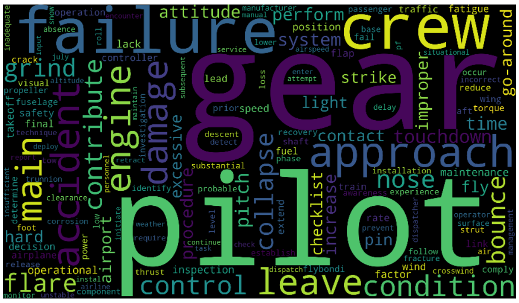


Figure 1. Wordcloud

The size of the word cloud font represents the frequency of these words in the accident investigation report, and the larger the font, the more times they appear. As can be seen from Figure 1, words such as gear(gear, landing gear), pilot(pilot), failure (loss of function, abnormal function of system equipment) appear frequently, which may be the main cause.

Next, the TF-IDF algorithm is used to extract text feature words and obtain the weight of each feature word to prevent important cause factors that are ignored due to low frequency.

2.2 TF-IDF weighting and feature word extraction

2.2.1 TF-IDF. The "word frequency-inverse document probability" method (TF-IDF) is a weighting technique commonly used in information retrieval[17]. This method was first proposed in 1988, when Salton and Bucley[18] proposed the probability TF-IDF statistical method of "word frequency - inverse document" to estimate the importance of a word in a document to a document in the corpus. If a word appears more frequently in a document but rarely in other texts, it is considered to have a good ability to identify the text, and this word is called a feature word[19].

Tf-idf is the product of TF and IDF, usually represented by w_t , as follows:

$$w_t = TF \times IDF \tag{1}$$

TF is the frequency of occurrence of the word t in the text, so the ability of the word to explain the text content can be expressed by calculating the value of TF, and the calculation formula is as follows:

$$TF = \frac{count(\omega)}{|D_i|} \tag{2}$$

count(ω) represents the frequency of keyword ω appearing in the text, and $|D_i|$ represents the total number of words contained in the text i.

IDF is the inverse document frequency, which represents the reciprocal of the occurrence times of the word w in the whole document set, describes the ability of the word to distinguish different documents, and reflects the importance of the word in the whole corpus from the perspective of information theory[20]. The formula is as follows:

$$IDF = \log \frac{N}{\sum_{i=1}^{N} I(w, D_i)} \tag{3}$$

N represents the total number of all documents, and $I(w, D_i)$ indicates whether the document i contains the keyword w, 1 if it does and 0 if it does not.

TF-IDF can effectively distinguish the feature words from the common words in the text. On the one hand, it can find out the important feature words that are ignored because of the low frequency of occurrence; on the other hand, it can filter out the common words that appear very frequently but are not practical for text analysis.

2.2.2 Feature word extraction. Since 181 accident investigation reports contain tens of thousands of words, the denominator (total number of text words) and numerator values are too different in weight calculation, resulting in the weight coefficient of extracted feature words being the same as seven decimal places, which makes it impossible to intuitively make a credible analysis of text data. Therefore, based on the original TF-IDF weighting algorithm, By splitting the text in equal amounts, a corpus with a little difference in the total number of hundreds of words is formed, the original parameter Settings are changed, and the weight coefficient of each feature word is obtained and the weight of the text can be intuitively seen. The weight coefficients of some feature words in a corpus are shown in Table 3 below:

Table 3. Part of the feature word weight

Number	Feature word	Weight coefficient
1	system	0.078101091
2	gear	0.058575818
3	failure	0.04144349
4	checklist	0.039050545
5	control	0.039050545
6	execution	0.039050545
7	aerodynamic	0.035765287
8	failure	0.035765287

2.3 Subject word mining

2.3.1 LDA model construction. LDA topic model is a typical bag of words model（Figure 2）, which regards the composition of a document as a collection of several words, and there is no sequential relationship between words.

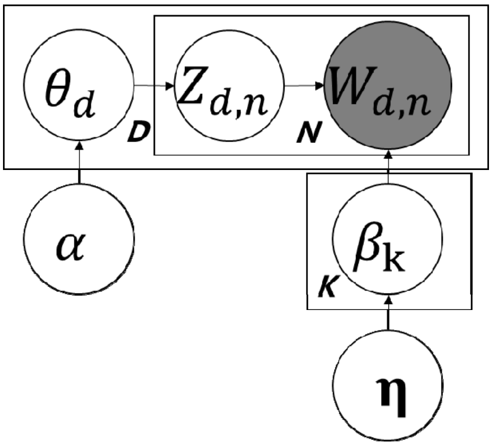


Figure 2. LDA

First, it is assumed that the prior distribution of accident investigation report subject follows Dirichlet distribution. For a document d, the distribution of its topics is:

$$\theta_d = Dirichlet(a) \quad (4)$$

The prior distribution of the words in the subject follows the Dirichlet distribution. Both for any subject k have:

$$\beta_k = Dirichlet\left(\eta\right) \tag{5}$$

For the NTH word in any accident report document d in the data, we can get the distribution of its topic number $Z_{d,n}$ from the topic distribution θ_d:

$$Z_{d,n} = multi\left(\theta_d\right) \tag{6}$$

Finally, for the topic number, we get the probability distribution of the word $W_{d,n}$ that we see:

$$W_{d,n} = multi\left(Z_{d,n}\right) \tag{7}$$

In this model, we have a Dirichlet distribution of J document topics, and the corresponding data has a multinomial distribution of J topic numbers, such that($\alpha\rightarrow\theta_d\rightarrow Z_{d,n}$)forms the Dirichlet-multi conjugacy, A posteriori distribution of the topic of a document based on the Dirichlet distribution can be obtained using the previously mentioned Bayesian inference method.

If the number of words for the KTH topic in the DTH document is $n_d^{(k)}$, then the count of the corresponding multinomial distribution can be expressed as $\overrightarrow{n_e} = \left(n_e^{(1)}, n_e^{(2)}, \dots n_e^{(k)}\right)$

Using Dirichlet-multi conjugation, the posterior distribution of θ_m is given as:

$$\theta_d = Dirichlet\left(\theta_e \middle| \vec{a} + \vec{n}_d\right) \tag{8}$$

In the same way, for the distribution of topics and words, we have the Dirichlet distribution of K topics and words, and the corresponding data has the multinomic distribution of K topic numbers, so that ($\eta\rightarrow\beta_k\rightarrow W_{d,n}$) forms the Dirichlet-multi conjugacy, A posteriori distribution of subject terms based on the Dirichlet distribution can be obtained using the previously mentioned Bayesian inference method.

Therefore, after deducing and simplifying the above formula, the mathematical formula of the LDA topic model constructed in this paper is formed:

$$p\left(w, z \middle| \alpha, \beta\right) = p\left(w \middle| z, \beta\right) p\left(z \middle| \alpha\right) = \prod_{k=1}^{K} \frac{\Delta\left(n_k + \beta\right)}{\Delta\left(\beta\right)} \prod_{m=1}^{M} \frac{\Delta\left(n_m + \alpha\right)}{\Delta\alpha} \tag{9}$$

Before constructing the LDA model, it is necessary to determine the number of topics K and hyperparameters, and ß, because the number of topics and the values of hyperparameters have a great impact on the results of the model. The value of ß is generally 0.1, and the value of min is 50/ number of topics. The number of LDA topic K is generally determined by consistency test or confusion degree calculation. This paper uses consistency test and confusion degree respectively to calculate the value interval of the number of text topics, and finally selects the value of the best topic number K by comparing the two.

The meaning of topic consistency is the similarity of the subject words under a certain topic in the corpus text. The consistency score under each topic can be obtained through the consistency calculation of the topic. The higher the consistency score, the better the number of topics. As shown in Figure 3, when the number of topics is in the range of 10 to 12, the consistency score is highest and the trend is flat.

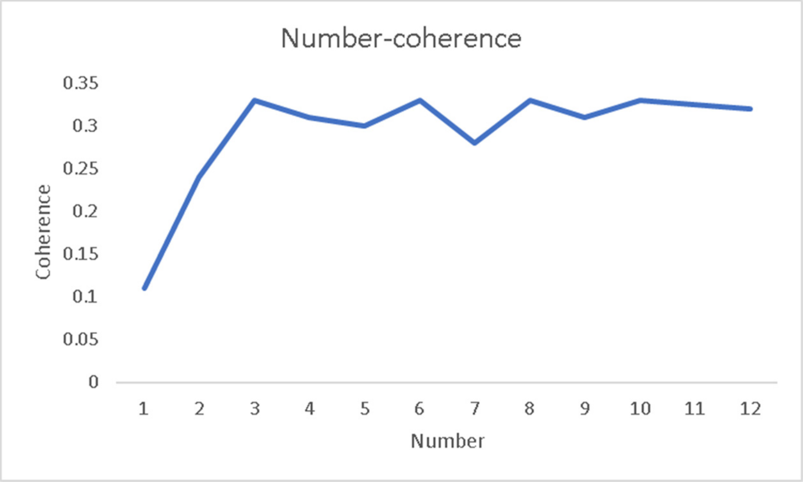


Figure 3. Consistency test

The meaning of the degree of confusion is the degree of confusion of the machine to the text. After repeated iterative operations, he concluded that the model ability and the degree of confusion are inversely proportional, that is, the smaller the degree of confusion, the better the effect of the model. In other words, the more topics there are, the less confusion there is. As shown in Figure 4 below:

$$K = Perplexity(d) = \exp\left\{\frac{-\sum_d \log\left(P\left(w_d\right)\right)}{\sum dN_d}\right\} \tag{10}$$

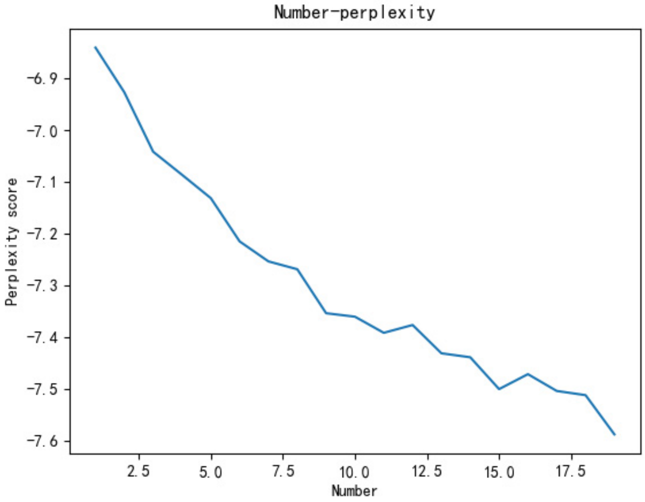
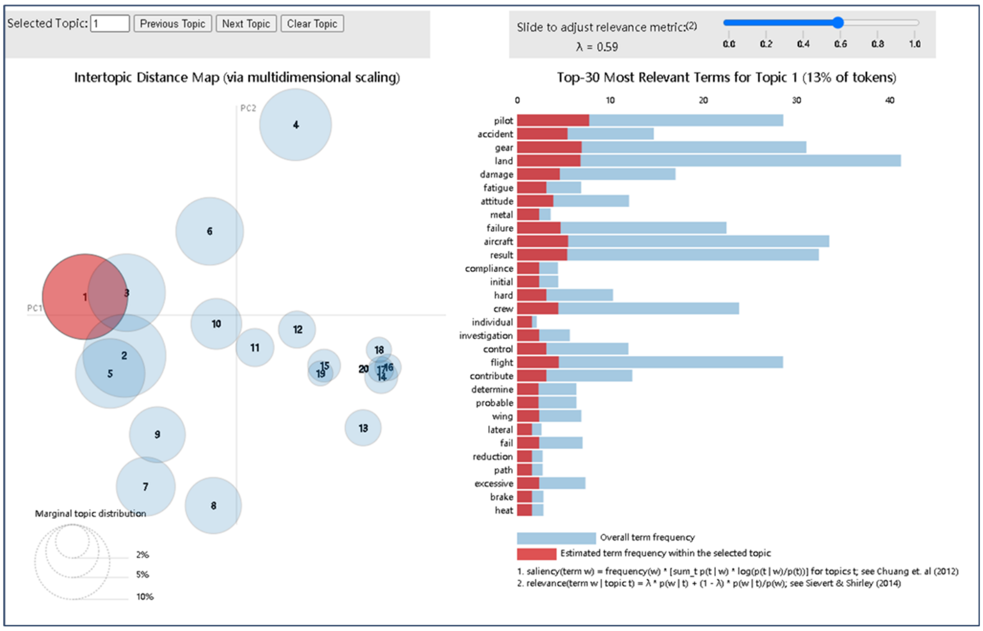


Figure 4. Confusion chart

Through consistency test and confusion degree calculation, it can be seen that when the number of topics ranges from 10 to 12, the consistency score is high, the topic confusion curve is low and the trend is stable. Therefore, the value of the number of topics K is selected from 10, 11 and 12. From theme-coherence we can see that the coherence score is higher when the number of topics is 10, while from the theme-confusion graph we can see that the coherence score is higher when the number of topics is 11 and 12. The machine is a little less confused by the text than when the value is 10.

Therefore, in order to better determine the value of K, experiments are conducted on LDA models with values of 10 and 11, and the interactive interface diagram of the LDA model is obtained as shown in Figure 5 below. When the number of topics is 11, it is found that topic 11 is almost included by topic 10, and thus cannot better represent the characteristics of this topic. When the number of themes is 10, it will be found that the independence of each theme is high, which can better represent the characteristics of each theme.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 5. LDA interactive interface diagram

Therefore, when the trainer pre-sets the number of topics in the training set, he will not select the number of topics corresponding to the minimum confusion value, but will select an interval in which the confusion degree continues to decrease but changes slowly, and then debug the number of topics by analyzing the effect of the theme words. Therefore, it is determined that the values of relevant parameters of the LDA model are set as lib =5, ß=0.1, K=10

2.3.2 Model solving. After solving the LDA model, the generated probability distributions of "subject-subject words" and "subject-document" are extracted, and the distribution of "subject-subject words" is conducted to organize the top 10 high-probability theme feature words corresponding to each theme, and the theme content is judged according to the high-probability theme feature words under each theme, as shown in Table 4 below:

Table 4. High probability topic feature words under each topic

Subject number	Top 10 high probability Thematic feature words
Topic 1	pilot,gear,damage,approach,main,failure, attitude,collapse,propeller,fuselage
Topic 2	failure, pilot, gear,factor,operation,tra, speed,approach,configure,instrument
Topic 3	poilt,condition,engine,performance,follow, lack,procedure,approache,operator,speed
Topic 4	gear,damage,pilot,contact,aft, main,system,trunnion,component,lead
Topic 5	pilot,monitor,procedure,flare,damage, gear,arm,single,consistent,prepare
Topic 6	approach,gear,collapse,low,controller, flap,main,short,flare,bounce
Topic 7	gear,pitch,condition,bounce,rate, touchdown,pilot,encounter,decision,attitude
Topic 8	failure,gear,main,body,event, strut,lower,system,pitch,attitude
Topic 9	night,detect,visual,inspection,flare, failure,pilot,engine,gear,failure
Topic 10	axis,IL,mlg,structure,correction, input,crosswind,thrust,technique,PF

From the output results of the LDA model, it can be seen that the ten themes divided and the representative feature words contained in each theme can be preliminarily speculated according to the correlation of these feature words. For example, the high-probability feature words "pilot", "landing gear", "operation method", "power system" and "failure" in theme 1 can be preliminatively speculated according to these feature words: (1) The wrong operation of the pilot leads to the failure of the landing gear, which is one of the important causes of runway accidents; (2) The failure or failure of the relevant power system is one of the important causes of runway accidents. After preliminary speculation on the content expressed by each topic according to the characteristic words of each topic, the topic probability distribution of each accident report text in the whole corpus is calculated to obtain the probability of each topic distribution of each accident investigation report.

The "subject-document" distribution obtained by solving the LDA model is shown in Table 5 below, which clearly presents the distribution probability of different topics in each accident document. The topic content in each accident document is found by the value. The larger the value, the greater the correlation

Table 5. Document topic distribution

Number	topic 1	topic 2	topic 3	topic 4	topic 5	topic 6	topic 7	topic 8	topic 9	topic 10	Subject
1	0.002858	0.002859	0.202971	0.002858	0.002859	0.002858	0.774162	0.002858	0.002859	0.002858	7
2	0.00526	0.00526	0.00526	0.00526	0.00526	0.00526	0.00526	0.00526	0.00526	0.95262	10

	4	4	4	4	4	4	4	4	4	6	
3	0.002632	0.002632	0.002632	0.002632	0.976313	0.002632	0.002632	0.002632	0.002632	0.002632	5
4	0.005883	0.005884	0.005883	0.947052	0.005883	0.005883	0.005883	0.005883	0.005883	0.005883	4
5	0.001924	0.001924	0.001924	0.001924	0.001924	0.726847	0.257764	0.001924	0.001924	0.001924	6
6	0.005264	0.952626	0.005264	0.005264	0.005264	0.005264	0.005264	0.005264	0.005264	0.005264	2
7	0.004546	0.004547	0.004546	0.004546	0.004546	0.004546	0.004546	0.004546	0.959084	0.004546	9
8	0.002084	0.002084	0.002084	0.002084	0.002084	0.002084	0.163407	0.819922	0.002084	0.002084	8
9	0.977497	0.002500	0.002500	0.002500	0.002500	0.002500	0.002500	0.002500	0.002500	0.002500	1
…	…	…	…	…	…	…	…	…	…	…	…
181	0.002942	0.002942	0.728338	0.002942	0.248124	0.002942	0.002942	0.002942	0.002942	0.002942	3

2.3.3 Topic analysis. The topic distribution probability of 181 accident investigation reports is calculated. On the one hand, the probability of each topic in a certain accident investigation report can be seen according to each row of the results in the above table, and the main topic of the accident investigation report can be found. On the other hand, according to each column of the results of the above table, the probability of a certain topic in different accident investigation reports can be obtained. Finally, the following conclusions are drawn from the analysis: In all the accident investigation reports, The probability of accidents caused by different themes was 16.5% for theme 1, 7.33% for theme 2, 17.2% for theme 3, 6.33% for theme 4, 2.47% for theme 6, 15.4% for theme 7, 4.33% for theme 8, 14.1% for theme 9, and 3.56% for theme 10.

3. Mining and analyzing association rules of Apriori algorithm

3.1 Construction of Apriori association rule model

First, the topic distribution of each corpus text is obtained through the LDA model and these topics are extracted. Secondly, the data set of the association rule model is constructed according to the format that each line represents each corpus topic. Finally, the association rule model is constructed according to the nature of the data set, as shown in Figure 6 below:

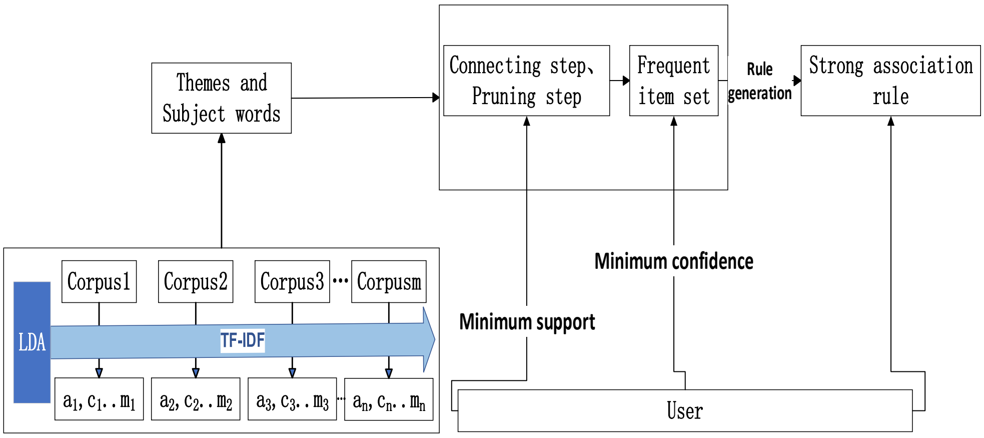


Figure 6. Schematic diagram of association rule model construction

3.2 The Apriori algorithm mines association rules

The Apriori algorithm was used to mine the airstrip accident investigation report to obtain 310 association rules, and 23 strong association rules were obtained under the conditions of "support > 20%, confidence > 80%, and enhancement > 1". Select 10 association rules with high confidence and large item sets, as shown in Table 6 below:

Table 6. Partially strong association rules

Number	Frequent items	Support degree	Confidence degree
1	land	0.805556	0.805556
2	gear,pilot,land	0.291667	0.966582
3	failure, pilot, land	0.263889	0.954857
4	collapse, gear, land	0.236111	0.932471
5	failure, gear, pilot,land	0.208333	0.931542
6	gear, main,land,	0.277778	0.921478
7	failure, gear, land	0.318213	0.914127
8	pilot,attitude,land	0.208333	0.891245
9	flare,land	0.215913	0.882353
10	approach,land	0.305556	0.846153

By analyzing the table, it can be concluded that there are some important factors that cause airstrip accidents: 1. The probability of airstrip accidents occurring in aircraft approach landing stage is much higher than that in other stages (Association Rules 1 and 10); 2. (2) Failure of the relevant part of the aircraft caused by the pilot's wrong behavior (Association Rules 2, 3, 5); 3. Landing gear collapse during landing (association Rule 4); 4. The power system of the landing gear fails during the landing stage (Association rules 6 and 7); 5. Working attitude of the pilot (Association Rule 8) 6. This series of operations in the landing phase.

3.3 Extraction of strong association rules

In order to more clearly and intuitively see the relationship between different accident factors and their interdependence, the association rules mined from the airstrip accident investigation report database based on Apriori algorithm in the previous section are visualized, and the generated association rules network diagram is shown in Figure 7 below.

The degree of interaction between each accident factor can be clearly seen through the association rule network diagram, and the cause of the accident can be analyzed by modifying the value range of the number of nodes. The data is mined and analyzed through the network diagram constructed by the strong association rules. The key nodes are selected according to the intensity of the association path of each node in the network diagram, and the key accident cause factors are found through the analysis of the key nodes. As can be seen from the figure below, the correlation paths of the following important nodes are dense and coupled with other nodes: {pilot, operational behavior, power energy system, system failure, landing gear, collapse, management system}.

Through the analysis of the path thickness of the network diagram and the strong association rules, we can see that: (1) there is a significant correlation between the airstrip accident and the pilot, system and management system. (2) The occurrence of airstrip accidents is often caused by the status or behavior of pilots. Combining the frequent terms (failure, gear, pilot,land) extracted in the above section and the corresponding association rules in the following figure, the cause chain "pilot status → operation error → relevant system failure → landing gear collapse → accident" can be analyzed. (3) The defect of the relevant management system is also an important reason for the occurrence of airstrip accidents. For example, the pilot's professional knowledge and relevant skills training are not perfect, which leads to the pilot's operation error, and then leads to the occurrence of accidents; The imperfect training of maintenance personnel in relevant aspects leads to errors in the inspection and maintenance of aircraft, which leads to the operation of aircraft with hidden trouble.

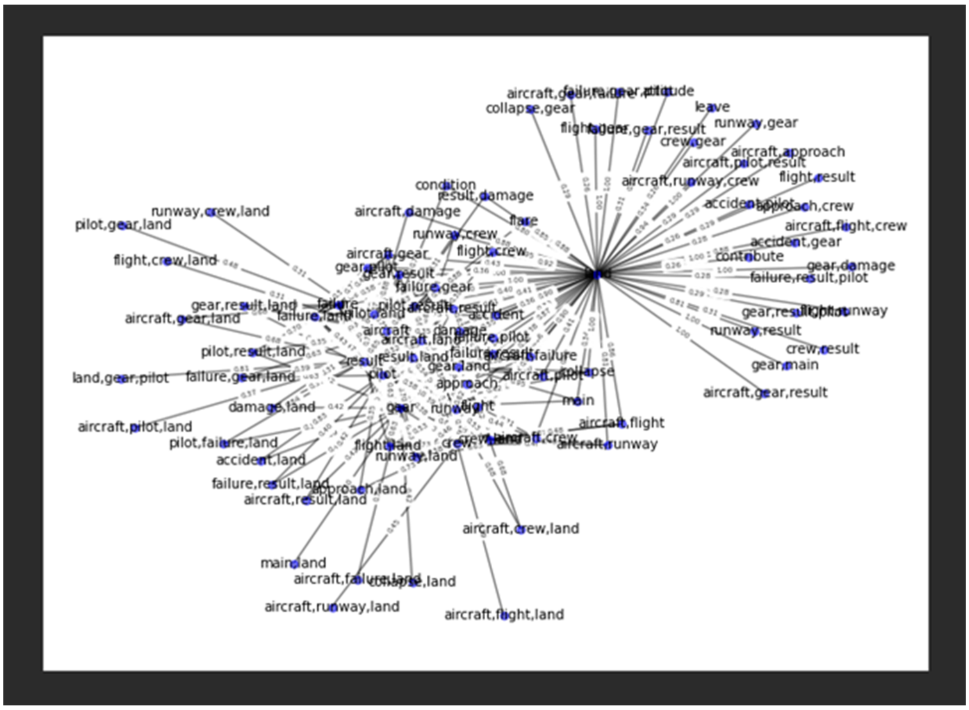


Figure 7. Association rules network diagram

4. Conclusion

1) Text mining technology was used to extract the ten accident-causing topics in the airstrip accident report and the high-probability subject words under each topic, and then the theme distribution probability was calculated for the extracted topics. Finally, it was found that the total probability distribution of theme 1, theme 3, theme 6, theme 8 and theme 9 was up to 80%. The extracted high-probability feature words with overlapping themes such as "pilot, operation method, state, power system, failure, landing gear, collapse, engine, system" are the key cause factors of airstrip accidents.

2) The Apriori algorithm is used to analyze the hidden association between the extracted topic and subject word, and the association rule network diagram is constructed. The key accident cause nodes and the association relationship between them are found based on the number and

thickness of paths of each critical accident node in the association rule network diagram.

3) The analysis results of network diagram and association rules show that: The five nodes, pilot status, pilot operation behavior, system failure, landing gear, and management system, have complex correlation paths and high correlation with other nodes, which are the key causes of airstrip accidents. Landing gear collapse is the most common type of airstrip accidents. The causation chain "pilot status → operation error → related system failure → landing gear collapse → accident" is extracted in this paper.

4) By collecting the investigation reports of airstrip accidents in the past 20 years for research and analysis, this paper finds the key accident types of airstrip accidents and their cause chain; Subsequently, the key accident types and cause chains in each time period can be studied by means of time line, and the evolution trend of runway accidents can be analyzed.

Acknowledgments: This work was financially supported by the Higher Education Talent Cultivation Quality and Teaching Reform Project of Sichuan Province,2021-2023(No.JG2021-316).

References

[1]Hong Chang. Application Research of New Aircraft Pavement Arresting System [D]. Civil Aviation University of China,2009.

[2]Zhiqin Huo, Yi Ru and Songchen Han. Analysis model of off-runway incident of Civil aviation transport aircraft during landing [J]. Journal of Southwest Jiaotong University,2012,47(05):895-900.

[3]Qinqian Lu and Tingting Yu. Anti-runway intrusion of multiple access runways [J]. Civil Aviation of China,2014(02):50-51.

[4]Xianli Zhao. Research on Evolution mechanism of airport runway safety risk [D]. Wuhan University of Technology,2019.

[5]Mengna Liu. Causative factor analysis and risk prediction of aviation safety accident report based on text mining [D]. Anhui University of Architecture and Architecture,2019.

[6]Xiuting Zhou. Research on risk control of Civil Aviation Batch unsafe events based on Text Mining technology [D]. Beijing University of Posts and Telecommunications,2021.

[7]Ma T. Visual analysis method of aviation safety accident report based on text mining and K-means clustering [J]. Computer Knowledge and Technology, 2002,18(35):56-59+72.

[8]Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3:993-1022.

[9]Griffiths T L,Steyvers M.Finding scientific topics[J].Proceedings of the NationalAcademy of Sciences of the United States of America,2004,101 Suppl 1(1):5228.

[10]Jingwei H,Yves L,Zbigniew K. Big Data Analysis with LDA for Cybersecurity in Organizations[J]. NATO Science for Peace and Security Series - D: Information and Communication Security,2017,48.

[11]Guimin H,Zhenglin S,Chunli F, et al. A Sentence Level Corrected LDA Topic Model for Chinese English Learners[J]. Frontiers in Artificial Intelligence and Applications,2019,314.

[12]Zhang C,Mei M,Mei Z, et al. PLDANet: Reasonable Combination of PCA and LDA Convolutional Networks[J]. INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL,2022,17(2).

[13]Jiazhi Sun. Research on Subject Discovery of Coal mine accident investigation report based on LDA [D]. Southwest Jiaotong University,2020.

[14]Lichao Yang, Jingxian Liu and Zhao Liu et al. Mining of causative association rules for water traffic accidents under value attenuation [J]. China Safety Science Journal,2022,32(10):

[15]Binbin Zheng, Tingting Feng and Jiahe Wang et al. Causes and correlation analysis of urban gas accidents based on text mining [J]. China Safety Science Journal,2023,33(07):190-195.

[16]Ling Yan. Model and the bayesian network based on themes of the road traffic accident cause research [D]. Lanzhou university, 2023. The DOI: 10.27204 /, dc nki. Glzhu. 2023.001979.

[17]Lei Zhang. Personalized push of hotel reviews based on word2vec and TF-IDF algorithm [J]. Computer and Information Technology,2017,25(06):8-11.

[18]G.Salton C B.Term weighting approaches in automatic text retrieval[J].InformationProcessing&Management,1987,24(5):513-523.

[19]Ponte J M, Croft W B. Text segmentation by topic[C]// European Conference on Research & Advanced Technology for Digital Libraries. 1997.

[20]Peixin Chen. Research on Vector representation and Modeling of text Semantics [D]. University of Science and Technology of China,2018.