Research on Enterprise Hidden Danger Association Rules Based on Text Analysis

Shengxin Ge, Yufeng Zhuang, Yanzhu Hu and Xinbo Ai*

School of Automation, Beijing University of Post and Telecommunications, Beijing 100876, China

*Corresponding author e-mail: axb@bupt.edu.cn

Abstract. Descriptive fields are used in the daily operation of an enterprise as a dimension for recording hidden danger letters. However, it is quite difficult to extract corresponding effective data from such text information to guide the operation of the enterprise. In view of this problem, this paper first grouped the hidden danger data of different enterprise types, then used Chinese word segmentation technology to transform the corresponding descriptive fields into structural data, extracted the topic model using Latent Dirichlet Allocation (LDA) algorithm, and finally used Apriori algorithm to find the association rules of hidden dangers under different enterprise types. After the analysis of the experimental results, the association rules validated accord with common sense and can provide data support for hidden danger supervision.

1. Introduction

How to ensure the safety of personnel in the production of enterprises is not only a corporate problem, but also a social issue. With the development of information technology, various organizations began to use electronic media to store daily operational data. These data also include information on production safety. It is a rather difficult problem that how to delete useless information from massive data and integrate effective information to obtain the required model to guide the operation of the organization. Especially for data such as production safety, most of them cannot be represented by data such as scalars, but texts are used to record data on production safety, such as production processes, hidden danger descriptions, and accident information. There is an urgent need to solve the problem of the complexity of semi-structured data in Chinese text, so that a large number of organizational safety production text data can be used in existing machine learning to obtain a model that can provide data support for future production safety. Therefore, research on information mining technology based on production safety text is very important and urgent.

Due to the rise of machine learning, corresponding algorithms have been applied to the data analysis of production safety management. In English, there is a study on the relationship between motor vehicle injury degree and hospital-related factors [1], analysis of fire factors in Spain [2], assessment of regional society, economy and environment through hydraulic fracturing activity [3], and study the relation between the full moon and traffic accidents through traffic accident information [4]. However, because of the limitations of some algorithm models, when text data is used in data analysis, most of the information is reduced. For example, the information of the text type data is
reduced according to the manually defined fault classification, resulting in the lack of fault detailed information in the data analysis. A small part of the use of word segmentation [5,6] to convert text information into a corresponding keyword network to achieve the introduction of text information. However, due to the different grammatical structures between Chinese and English, the corresponding results can not be directly applied to the analysis of Chinese texts.

With the development of Chinese words segmentation, such as text segmentation technology based on analysis [7], Context Information and Fragments Based Cross-Domain Chinese Word Segmentation [8], and A new feature selection method for handling redundant information in text classification [9]. How to make full use of the text data that has been reduced or even neglected in the past, the detailed data from the cause to the result is completely incorporated into the data analysis, and finally the model that can be used completely and comprehensively for event correlation analysis begins to be paid more and more attention by researchers. Now there is an analysis of the comprehensive strength of tourism in Zhejiang Province [10], using text mining rheumatoid arthritis and diabetes to understand the mechanism of traditional Chinese medicine treatment of different diseases [11], power work sheet text data analysis mining model research [12] and other articles The application of text mining has achieved good results. However, they are all a classification study on one aspect. The final result is the relationship between cause and effect, without taking into account the relationship between effect and effect. In daily safety production management, it is necessary not only to prevent accidents caused by corresponding hidden dangers, but also to consider the problems of concurrent accidents caused by the connection between them. This article is about research in this area.

In this paper, data analysis is carried out based on The R Programming language. The first step is to use the Chinese word segmentation Rwordseg in the R environment to perform word segmentation and word frequency statistics on the description fields and the sub-category fields of the hidden danger information. The second step is to classify the enterprise information according to the types of regulatory industries, and to extract and classify the data after the segmentation. The third step is to further analyze the relationship between the types of enterprise supervision and the types of hidden dangers, to obtain the potential risks of the identified types of enterprises, and to obtain the relationship between the hidden dangers of a certain type of enterprises.

2. Experimental Design
The first step of the experiment is to preprocess the data, which is mainly divided into two parts. First, extract the description text in the hidden danger information record without missing and can be used normally. Use the Chinese word segmentation Rwordseg in the R environment to perform Chinese word segmentation on the corresponding description text. Delete the useless information of the word frequency data information after the word segmentation, such as auxiliary words, prepositions and the like. This package uses rjava to call Java's Chinese word segmentation tool Ansj, which is based on the icetlas Chinese word segmentation algorithm of the Chinese academy of sciences and adopts Hidden Markov Model (HMM). The algorithm implements the following steps: 1. Full-segmentation, atomic segmentation; 2. N-short path rough segmentation, according to hidden Markov model and viterbi algorithm, to achieve optimal path planning; 3. Name recognition; 4, system dictionary supplement; 5, user-defined dictionary supplement; 6, part-of-speech tagging (optional). When using this package for Chinese word segmentation of description field of hidden danger, some professional words cannot be distinguished correctly, so professional words can be distinguished correctly by adding dictionary. After extracting the useful data, the result of word segmentation is transformed into the word frequency matrix. Second, extract the account information of the corresponding enterprise according to the record extracted from the first.

The second step is to classify companies and hidden dangers, which are also divided into two parts. First, classify enterprises according to the types of regulated industries. Secondly, the classification of hidden danger is to classify hidden danger topics based on LDA topic model algorithm and extract the topics for a single record. Compared with the commonly used clustering algorithm, which regards a
record as a unit, LDA thematic model algorithm assumes that each record is composed of multiple different topics in different proportions, making the classification more detailed. In this way, the text information that was not used in the big data analysis of the past hidden dangers is fully utilized, so that the classification obtained is more detailed than in the past. This makes the final analysis more applicable to the actual situation.

The third step is to find out the pattern of hidden dangers in different types of enterprises. According to the classification in a single enterprise, the number of hidden dangers is converted into the proportion corresponding. Synthesize the hidden dangers of each enterprise under the single enterprise classification, and finally use the Apriori algorithm to obtain the association rules between them.

After the above three steps, an enterprise can be classified. After the classification, other potential risks can be predicted through the detected hidden dangers to provide data support for the hidden danger investigation work.

Figure 1. Experiment process

3. Experiment Procedure

3.1. Data Preprocessing
Clean the raw data. The first step is to screen out the enterprise account records for which the required regulatory industry type fields are not missing. In order to avoid the lack of information in the hidden danger record due to recording errors or irregularities. Eventually, the analysis results in error. Remove the bad record that the description field is empty or the related description field is less than 10 words after word segmentation. Through the above methods, the inferior record can be largely removed, and finally the processed hidden danger record table is obtained. Then extract the company name from the table. Corresponding to the previously extracted enterprise account information record, the complete account information table and the hidden danger information table are obtained. Finally, there are 212,608 hidden danger records of 1964 companies. Among them, there are 21 types of enterprise supervision.

The corresponding experimental process and results are briefly described by taking industrial-mechanical enterprises as examples.
The names of the enterprises in the industrial-mechanical classification are extracted from the enterprise account information table processed above, with a total of 172 enterprises. According to the enterprise name, the corresponding records are extracted from the processed hidden danger information table. There are 14088 records in total.

Extracting the information in the text requires that the text be first converted to the corresponding word frequency vector for subsequent operations. In this paper, the segmentationCN function based on the Rwordseg package in the R language is used to process the corresponding sentences and finally convert them into word frequency vectors. The word vector that combines all hidden trouble records becomes the word frequency vector matrix, and each row represents a word frequency vector corresponding to a record.

3.2. Classification

In the LDA topic model algorithm, it is assumed that the data set contains a total of K topics, and each piece of data is composed of different topics in different proportions. It is necessary to determine the appropriate number of topics so that the classification between the finalized topics is clear and the distribution of the topics of the data is realistic.

The calculated topic model Perplexity is usually used to determine the optimal number of topics. Its formula is as follows:

\[
\text{perplexity}(D_{\text{test}}) = \exp\left\{ \frac{-\sum_{d=1}^{M} \log(p(w_d))}{\sum_{d=1}^{M} N_d} \right\}
\]

(1)

\[
p(w_d) = \sum_{z} p(z) \cdot p(w|z, \text{gamma})
\]

(2)

Where M is the size of the test corpus, N_d is the text size (i.e. number of words) of part d, z is the topic, w is the document, and gamma is the text-topic distribution learned from the training set. When the number of topics is calculated from 2 to 100, the model perplexity changes. The results are shown in Figure 2 below:

![Figure 2. Topic model Perplexity](image)

The abscissa is the topic number and the ordinate is the model perplexity. As shown in the figure, it is necessary to select the K value at the position where the model perplexity is relatively flat, but the number of topics thus obtained tends to be too large, resulting in the semantic ambiguity of the topic. The essence of the topic extraction model is to divide the document into detailed and clear topics. Therefore, it is very important to consider the difference between the topics. This paper uses the mean square error between the calculated topics to judge the difference between the topics. The formula is as follows:
Where $K$ is the number of topics and $q_n$ is the word probability vector of the corresponding topic. When the number of topics is changed from 2 to 100, the mean square error between the topics changes. The result is shown in Figure 3 below:

$$D = \frac{\sum_{n=1}^{K} (q_n - \sum_{n=1}^{K} q_n)^2}{K}$$  \hspace{1cm} (3)$$

The abscissa is the topic number and the ordinate is the degree of confusion. It can be seen from the figure that, starting from the number of topics is 15, the increase speed of the mean variance of topics becomes slow, that is, the change degree of differences becomes smaller. Therefore, the clarity of confusion, difference and semantics should be taken into comprehensive consideration, and 20 should be selected as the appropriate number for the subsequent topic analysis.

Finally, the distribution of the top ten words of the 20 hidden danger topics obtained from the 172 companies from the industrial mechanical type through the LDA topic model algorithm is shown in Table 1 below:
Table 1. Top ten words of each topic

| Theme 1 | Theme 2 | Theme 3 | Theme 4 | Theme 5 |
|---------|---------|---------|---------|---------|
| registration, Special equipment, exist, scrap formalities, To handle the, To achieve, transform, The value of, certificate | practitioners, homework, supplies, Not allowed to, operation, Operating procedures, prohibited, achieve, work, Labor protection | device, To protect the, emergency, Safety protection, electrical, broken, automatic, brake, instructions, alarm | goods, Forklift truck, evacuation, The handle, More than, number of people, secure, The direction of, open, driving | The power supply, Fire control facilities, happen, Shut down, equipment, stop, test, accident, control, Cut off the |
| Theme 6 | Theme 7 | Theme 8 | Theme 9 | Theme 10 |
| protective, device, installation, parts, fixed, pulley, The belt, occur, movement, The shaft | office, place, production, clean, health, The ground, general, Clean and tidy, items, corridor | channel, fire, Take up, Shall not be, The power distribution, The workshop, Secure channel, neat, A security exit, store | basis, turn, fire truck, screening, organization, accidents, Less than, meet, The width of the, clearance | The surface of the, equipment, burr, protruding, Edges and corners, Is greater than, mm, rotating, shield, Parts of the |
| Theme 11 | Theme 12 | Theme 13 | Theme 14 | Theme 15 |
| wear, operation, Equipped with, Protective equipment, equipment, practitioners, Enter the, To adapt to, aging, Individual protection | rotating, movement, Fire extinguisher, between, unit, parts, The water tank, At the top of the, The sample, mm | Whether or not, running, operation, check, objects, touch, across, transfer, good, dangerous | The ground, equipment, Can't, store, Oil pollution, clutter, The production line, check, The surrounding, health | device, check, shield, Need to be, Power outages, Remove the, equipment, To turn it off, Adjust the, Machines and tools |
| Theme 16 | Theme 17 | Theme 18 | Theme 19 | Theme 20 |
| items, check, The machine, may, Parts of the, screening, running, maintenance, carry out, parking | Whether or not, Set up the, In good condition, mark, positioning, screw, The buffer, Hammer head, The mould, loose | Labor protection articles, switch, use, wear, requirements, Conform to the, correct, countries, The industry standard, practitioners | work, equipment, boot, operation, personnel, use, Engaged in the, things, leave, Irrelevant | happen, Electrical appliances, The fault, equipment, personnel, device, On the, Flammable items, signal, demolition |

3.3. Association Rules Mining and Results Analysis

After obtaining the distribution of the topics for each hidden danger record. According to the goal of the text-based analysis of enterprise hidden danger association rules in this paper, it is necessary to consider the distribution of the topic of each hidden danger record under the same enterprise. Considering that the association rules between hidden dangers do not involve the discussion of a single, and the data needs to be converted into discrete data in the Apriori algorithm, it is stipulated
that the proportion of topics with a single topic ratio of less than or equal to 0.1 in a record is set to 0, and the value of 0.1 or more is set to 1. When considering the hidden dangers of an enterprise, it is necessary to consider the time relationship between them. So it is necessary to merge the records over a period of time. The vector of the same time period after processing is added by OR operation. Finally, for each enterprise, form a sum vector of one or several time periods. Finally, 14088 hidden trouble dangers of 172 enterprises were transformed into 603 vectors about 20 topic classification, and then the association rules were excavated.

Finally, rules 5242201 are obtained. The number of rules increases with the increase of LHS. The changes are as Table 2 follows:

| Number of LHS | Number of rules |
|---------------|-----------------|
| 1             | 22              |
| 2             | 307             |
| 3             | 1922            |
| 4             | 7751            |
| 5             | 23256           |
| 6             | 54264           |
| 7             | 100776          |
| 8             | 151164          |
| 9             | 184756          |
| SUM           | 524220          |

As the length of the LHS increases, the number of rules increases. There are only 223 rules when there is only one LHS, and when the number of LHSs reaches nine, the relevant rules reach 1847560. But most of these rules are not valuable rules, that is, they do not provide much help in hidden danger investigation. The following is a discussion of selecting some valuable rules by lift.

Under the different 10 RHSs, the ten rules for ranking first by lift are as Table 3 follows:

| LHS                              | RHS | support     | confidence | lift       |
|----------------------------------|-----|-------------|------------|------------|
| 5,7,10,13,15,19                   | 1   | 0.06135987  | 0.9487179  | 4.931698   |
| 1,3,12,14,20                      | 2   | 0.03980100  | 1.0000000  | 2.491736   |
| 4,5,8,11,13,19                    | 3   | 0.05638474  | 0.9444444  | 4.126812   |
| 7,10,12,15,19                     | 4   | 0.06135987  | 0.9736842  | 3.208369   |
| 12,13,19,20                       | 5   | 0.04643449  | 1.0000000  | 2.441296   |
| 12,16,17,19,20                    | 6   | 0.04145937  | 1.0000000  | 4.568182   |
| 1,4,12,17,19                      | 7   | 0.04477612  | 1.0000000  | 1.866873   |
| 2,5,6,7                           | 8   | 0.09121061  | 0.9821429  | 1.961034   |
| 11,12,17,19                       | 9   | 0.04809297  | 1.0000000  | 2.451220   |
| 6,12,20                           | 10  | 0.06135987  | 1.0000000  | 2.599138   |

Among them, the four rules of 1, 3, 4, and 6 have a degree of lift greater than 3, which is considered to be a rule in which the rule of value is applicable in most cases. For example, in the first rule, in the case of the hidden danger classification topic 5, 7, 10, 13, 15, 19, the confidence and lift of the topic 1 are quite high. Indicates that the correlation between LHS and RHS is extremely high. All these classification topics in LHS need to be recorded, so the hidden danger in these aspects must be caused by the staff’s failure to timely find out the problems in recording and registration. The theme 1 corresponds to the hidden dangers of registration. According to the above analysis, the association rules proposed in this paper are consistent with common sense, and thus are applicable to the investigation of hidden dangers.

The relationship network diagram of the rules in Table 2 is as Figure 4 follows:
The starting point of the arrow is LHS, the ending point of the arrow is RHS, the number represents the type of topic, and the size of the circle in the middle indicates the relative confidence of this rule. It can be seen from the figure that the topic 19 appears most frequently in the LHS segment, indicating that the topic 19 is a fundamental hazard in the industrial-mechanical type of enterprise, and most of the other hidden dangers are related to it. Theme 19 is the specification problem of the operator when operating the machine. The industrial-mechanical type enterprise focuses on the mechanical operation, so the irregular operation of the machine will lead to a series of other hidden dangers. This on the other hand shows the correctness of the resulting rules.

4. Conclusion
Use Chinese word segmentation to convert text into word frequency vectors. The LDA algorithm that determines the optimal number of topics by the topic model Perplexity, topic difference and topic clarity. The descriptive text of the hidden danger information is transformed into a suitable and appropriate distribution of topics. This kind of semi-structured data is transformed into structured data with clear classification. The Apriori association mining algorithm is applied to the transformed structured data to obtain the association rules between hidden dangers.

The descriptive texts of the hidden dangers in the research are semi-structured texts, from which the association rules between various types of hidden dangers under different types of enterprises are mined. It provides data support for guiding enterprises' self-inspection of hidden dangers in daily management and relevant law enforcement departments to focus on different types of enterprises' hidden dangers.

Acknowledgments
This work was financially supported by Beijing Municipal Science and Technology Project (No. Z181100009018003).

References
[1] Cirera E, Plasència A, Ferrando J, et al. Factors associated with severity and hospital admission of motor-vehicle injury cases in a southern European urban area [J]. European Journal of Epidemiology, 2001, 17 (3):201-208.
[2] Rodrigues M, Jiménez-Ruano A, Peña-Angulo D, et al. A comprehensive spatial-temporal analysis of driving factors of human-caused wildfires in Spain using Geographically Weighted Logistic Regression [J]. Journal of environmental management, 2018, 225: 177-
[3] Mohtar R H, Shafiezadeh H, Blake J, et al. Economic, social, and environmental evaluation of energy development in the Eagle Ford shale play [J]. Science of The Total Environment, 2019, 646: 1601-1614.

[4] Onozuka D, Nishimura K, Hagihara A. Full moon and traffic accident-related emergency ambulance transport: A nationwide case-crossover study [J]. Science of the Total Environment, 2018, 644: 801-805.

[5] Suryani A A, Widiantoro D H, Purwarianti A, et al. The Rule-Based Sundanese Stemmer [J]. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 2018, 17 (4): 27.

[6] Šajgalík M, Barla M, Bieliková M. Exploring multidimensional continuous feature space to extract relevant words [C]//International Conference on Statistical Language and Speech Processing. Springer, Cham, 2014: 159-170.

[7] Šajgalík M, Barla M, Bieliková M. Exploring Multidimensional Continuous Feature Space to Extract Relevant Words [M]// Statistical Language and Speech Processing. Springer International Publishing, 2014: 159-170.

[8] LIU M, WANG X, LIU Y. Research on Text Segmentation Based on Topic Analysis [J] [J]. Acta Electronica Sinica, 2009, 2: 007.

[9] Degen H, Deqin T. Context information and fragments based cross-domain word segmentation [J]. China Communications, 2012, 9 (3): 49-57.

[10] Wang Y, Feng L. A new feature selection method for handling redundant information in text classification [J]. Frontiers of Information Technology & Electronic Engineering, 2018, 19 (2): 221-234.

[11] Guanghai Z, He G. An Analysis of the Spatial-Temporal Evolution of Urban Tourism Comprehensive Competitiveness in Zhejiang Province [J]. Journal of Ocean University of China (Social Sciences), 2015, 1: 013.

[12] Zhao N, Zheng G, Li J, et al. Text mining of rheumatoid arthritis and diabetes mellitus to understand the mechanisms of Chinese medicine in different diseases with same treatment [J]. Chinese journal of integrative medicine, 2018: 1-8.

[13] Zhu J, Cheng Y, M. Research on text mining model of electric power work [J], Power Demand Side Management, 2017, 19 (S1):87-89.