Application of Text Mining Techniques in Railway Safety Supervision System

X Q Li¹, T Y Shi², P Li³, F Gao¹ and W G Xiang¹

¹Postgraduate Department, China Academy of Railway Sciences, Beijing 100081, China
²Institute of Computing Technologies, China Academy of Railway Sciences, Beijing 100081, China
³Research and Application Innovation Centre for Big Data Technology in Railway, China Academy of Railway Sciences, Beijing 100081, China
lixin123@sina.com

Abstract. Railway safety is the most important in railway transportation. It is the cornerstone of all railway work and involves a wide range of departments. In the age of unstructured big data, it is an urgent problem to solve unstructured data, which contains a lot of information value by using advanced technology. This paper introduces the application of text analysis in railway safety supervision system based on massive data of accident and fault analysis report, analysis the importance of text mining techniques for railway safety system, the key technology of text analysis. Finally, based on the Railway Security Big Data Application Platform, we mainly introduce two classic functions of full-text retrieval and text classification based on text analysis technology on this platform. The application of the platform proves that text analysis plays a key role in railway safety supervision system.

1. Railway Safety System

With the development of railway informatization, the railway has used to update the advanced information technology, real-time monitoring the affecting the railway safety of people, equipment, environment and other state through various professional monitoring system and information sharing technology, to real-time find problems and solve the problem [1]. Management personnel also regularly go to the site inspection, record the problems found, analysis the risks and hidden dangers caused by the problems, and issue the safety notice in real time. After an accident or fault happen, fill in the event details and submit them to the business offices for processing according to the procedures. In the process of accident or fault treatment, the event details, analysis information and treatment measures are constantly improved, and finally the accident fault analysis report is formed and archived.

Railway accident and fault analysis report mainly records the detail of accident and fault, including information such as time, place, train, and the analysis of the cause of the accident or fault, lessons learned and corrective actions, and the analysis of the responsibility, finally do qualitative rules for the accident or fault. For a long time, from these, the managers to master the situation, reason and rule of accident and fault, aim at key management to the weak of safety production, take specific measures to avoid accident and fault. Implementing countermeasures to serve the railway transportation production system is the main basis for the railway security department to improve the safety system. As Figure 1 shows, but this feedback management pattern is often due to human analysis of text data, exits some
problems such as incomplete analysis and untimely analysis. How to apply text analysis technology to assist personnel to analyze the data automatically and effectively improve work efficiency and quality is an urgent problem to be solved by railway safety supervision department, so text mining techniques like text classification, text clustering, cause recommended, etc, are used to analysis accident and fault analysis report is an effective way.

![Safe production management model](image)

Figure 1. Safe production management model

2. Text Mining Techniques
Text analysis is a technical term that has arisen again with artificial intelligence in recent years, as well as a pattern or knowledge analysis technique that extracts previously unknown, computer-comprehensible, and of practical value from a large number of unstructured texts [2]. The text analysis process consists of text preprocessing, text feature extraction and knowledge discovery. Text preprocessing is realized by the basic techniques of text segmentation, including word segmentation, word tagging, and syntactic analysis. Feature extraction based on the traditional model by BOW bag model representation, TF-IDF word frequency and theme LDA model. Analytical model includes deep learning model of convolution based neural network (CNN), circulation neural network (RNN), and long-short term memory (LSTM), etc. Knowledge discovery process is the implementation of application scenarios, typical text analysis application scenario includes information extraction, such as keywords, the named entity extraction of text, the text label of automatic classification and automatic clustering without labels, based on the storage and retrieval of full text retrieval functions, full text summarization, and application of knowledge map of diversified technology, based on knowledge map of intelligence Q&A, as Figure 2 shows.
Figure 2. Text analysis techniques, models, and typical application scenarios

3. Typical application of text analysis in Railway Safety Big Data Platform

Railway Safety Big Data is a security big data application platform integrating massive data integration, sharing, storage, analysis and visualization of railway security monitoring information systems with big data technology, the overall architecture is shown in figure 3. The platform is divided into data storage, data analysis and application scenarios. Data storage platform with every business system of railway safety data as data sources, providing structured data and unstructured data of different data types, such as data access ability, the original data from the data source layer, through the extraction, conversion, standardization, then provide data to sharing platform. Data analysis platform provides real-time monitoring and analysis of massive heterogeneous data, including statistical analysis, multidimensional analysis, mining algorithm database, data mining tools, simulates and forecasts massive data. The application scenario of railway security supervision is an application platform that assists personnel in analysis and mining data with interactive interface according to data characteristics and application requirements, and presents data in different dimensions.

Figure 3. Overall architecture of secure big data platform
Based on the Railway Safety Big Data Application Platform, this paper mainly introduces the storage and full-text retrieval function of the railway accident fault analysis report, and take electrical engineering as an example, introduce typical application of text classification for automatic classification of accident and fault analysis report. On the platform, the application effect is verified by taking more than 1000 accident failure analysis reports of Taiyuan railway bureau from July 2016 to December 2017 as an example.

3.1. Full-text retrieval technology

In this paper, ElasticSearch (ES), a distributed full-text retrieval engine based on Lucene, is used to store and retrieve text data. ES is a database that uses inverted indexes and json as a distributed storage of unstructured data in document serialization format, compared with relational database, its retrieval speed has a significant advantage.

3.1.1. Data Storage

ElasticSearch uses Index=> Type=>Segment=>Document=>Field=>Term to store data [3], where an index refers to a logical namespace that contains several types. A document type is the concept of a table in a relational database, it is a grouping of indexed information that contains information about segments. A document is like a row of data in a relational database containing several fields. Domain information including the domain name and domain values, the domain name is a string, the threshold value is an item, the index is the smallest unit, represents a keyword and its in the original position, the number of occurrences of such information.

3.1.2. Data retrieval

Full-text retrieval uses the user input search keywords, according to the keyword query most relevant documents. In this paper, we use TF-IDF algorithm to realize retrieval, TF-IDF is a weighted technology based on information retrieval, the main idea is: if the search keyword appears frequently in one document with high TF and rarely in other documents, it is considered that the word or phrase has good discrimination and is suitable for classification [4].

Term frequency (TF) refers to the number of times the search keyword appears in the field text, and the more times it appears, the more relevant the search is to the document. Inver document frequency (IDF) indicates the number of times the search keyword appears in all documents, and the more it appears, the less relevant it is to these documents.

In a document, word frequency refers to the frequency at which a given word appears in the document, and is a normalization of the number of occurrences. For the word $t_i$, importance in a document can be expressed as Equation 1.

$$t_{f_{i,j}} = \frac{n_{i,j}}{\sum_{i} n_{i,j}}$$

In the above formula, $n_{i,j}$ is the number of times the word $t_i$ in document $d_j$ appears, the denominator is the sum of all the words in $d_j$.

Reverse file frequency is a measure of the general importance of a word, which is divided by the total file data divided by the data of the file containing the word, and the result is logarithmic, as shown Equation 2.

$$idf_i = \log \frac{|D|}{|j : t_i \in d_j|}$$

$|D|$ is the total number of files, $|j : t_i \in d_j|$ is the number of files containing the word.

then $tfidf_{i,j} = t_{f_{i,j}} \times idf_i$, the high weight TF-IDF is obtained by multiplying the high word frequency in a document by the low file frequency of the word throughout the document set.
In this paper, the accident analysis report of Taiyuan railway bureau is stored in ES database. Java technology based UI and backend ES storage and retrieval technology are combined to realize full-text retrieval. The practice proves that with the increasing number of documents, the retrieval speed is not affected.

3.2. Text classification

Text classification is one of the classic applications in natural language processing, its main purpose is to automatically classify massive data into predefined categories, which is an effective means to manage and utilize massive text information [5]. This paper classifies the accident report and analysis text of railway electrical engineering major as an example.

We choose the LSTM (Long Short-Term Memory) model to realize the text classification, the LSTM is comprised by embedding layer and LSTM layer [6]. We use character embeddings layer to construct character representation for each Chinese character in the input sentence, then feed them into the LSTM layer, the vector used here is word2vec in the Chinese wiki corpus which using genism [7]. LSTM is a special type of RNN (Recurrent Neural Networks), it can learn long-term dependency information, for particular construction to interact information, for it has three gates to protect and control cell state, they are forget gate, input gate and output gate, as Figure 4 shows.

![Figure 4. LSTM special structure](image)

Forget gate decides what information do we throw from the cell state; Input gate determines what new information is stored in the cell state; Output gate determines what the output is, the specific workflow of LSTM unit is expressed by the following equations.

\[ i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \]  
\[ f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \]  
\[ o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \]  
\[ C_t = f_t \cdot C_{t-1} + i_t \cdot \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \]  
\[ h_t = o_t \cdot \tanh(C_t) \]

Among them, \( i_t \), \( f_t \), \( o_t \), \( C_t \) respectively represent \( t \) time input gate, forget gate, output gate and cellular status output, \( x_t \) and \( h_t \) represent input vector and hidden layer vector at time \( t \), \( \sigma \) indicates that the sigmoid activation function, \( W_i \) and \( b_i \) respectively represent the weight matrix and the offset matrices, and their subscripts represent their respective values.

The experiment result as Table 1 show, we may easily find the CTC, TDCS, turnout, locomotive and interlock text have high accuracy rate and recall rate. This is because this type of texts
has good textual features. Filament power supply panel and Light band texts record diversification. At the present stage, classification results need to be corrected manually.

| Text type                  | Accuracy rate | Recall rate | F1    |
|----------------------------|---------------|-------------|-------|
| CTC                        | 94.23%        | 91.13%      | 92.65%|
| TDCS                       | 93.15%        | 90.45%      | 91.78%|
| Turnout                    | 96.56%        | 95.67%      | 96.11%|
| Filament power supply panel| 89.34%        | 88.24%      | 88.79%|
| Light band                 | 82.23%        | 81.02%      | 81.62%|
| Locomotive                 | 93.34%        | 92.48%      | 92.91%|
| Interlock                  | 95.23%        | 94.63%      | 94.93%|

4. Conclusion
The application of railway text analysis based on the Railway Security Big Data Platform is a breakthrough application for railway, which solves the barrier of railway text unstructured data which cannot be processed by computer for a long time. However, there are many problems in railway application scenarios and text analysis techniques. First, existing applications are not a high accurate, taking text classification as an example, the effect of text classification needs to be corrected by human, and the accuracy of text classification needs to be improved. Second, single application scenario, on the railway safety big data platform, it mainly realizes functions such as full-text retrieval, text classification, accident correlation analysis and named entity extraction, however, based on the existing unstructured text data, more dimension analysis needs to be dug to present the data. Third, key technologies need a breakthrough, with the advent of the era of artificial intelligence, which is given priority to with the depth study of breakthrough and the application of the technology has been widely, railway needs to absorb advanced technology, break through the existing technical barriers, improve application effects. Last, lack of railway knowledge database [8], the technology of word segmentation and knowledge map based on professional dictionaries all require a comprehensive professional corpus and knowledge database, therefore, the construction of railway professional corpus and knowledge database is conducive to the long-term development of railway text analysis.

Acknowledgments
financial support: China Academy of Railway Sciences within the major issues of the fund (2017Y3005)

References
[1] Wang L 2017 Analysis of the Importance of Process Control in Railway Safety Management Engineering technology: citation 2 00119-119.
[2] Dou Z C 2015 Opportunities and challenges of text big data analysis technology Financial Computerizing 11 59-61
[3] Zhang J Z, Huang Y F, Xiong Y J 2015 Digital Library R retrieval System Based on ElasticSearch Jisuanji yu xiandaihua 6 69-73.
[4] Yang L B, Li P 2018 Intelligent Classification of Fault of Railway Signal Equipment Based on Imbalanced Text Data Mining Journal of the china railway society 2 59-66.
[5] Ahmad K, Vrusias B, Oliveira P C F D 2003 Summary evaluation and text categorization ACM 443-444.
[6] Zhou C, Sun C and Liu Z A 2015 C-LSTM Neural Network for Text Classification Computer Science 4 39-44.
[7] Řehůřek, Radim and Sojka, P 2010 Software Framework for Topic Modelling with Large Corpora 45-50

[8] Hsu M J, Ho C P 2012 Creating a knowledge discovery model using MOEX's examination database for in-depth analysis and reporting Robotics and Applications. IEEE 705-707.