energies

MDPI

Article

A Network Method for Identifying the Root Cause of High-Speed Rail Faults Based on Text Data

Liu Yang [1], Keping Li [1,*], Dan Zhao [2], Shuang Gu [1] and Dongyang Yan [1]

[1] State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China; lyang90@bjtu.edu.cn (L.Y.); 17114248@bjtu.edu.cn (S.G.); 18114013@bjtu.edu.cn (D.Y.)

[2] Key Laboratory of Data Engineering and Knowledge Engineering of Ministry of Education, School of Information, Renmin University of China, Beijing 100872, China; cdanzhao@ruc.edu.cn

* Correspondence: kpli@bjtu.edu.cn

Received: 21 April 2019; Accepted: 14 May 2019; Published: 18 May 2019

Abstract: Root cause identification is an important task in providing prompt assistance for diagnosis, security monitoring and guidance for specific routine maintenance measures in the field of railway transportation. However, most of the methods addressing rail faults are based on state detection, which involves structured data. Manual cause identification from railway equipment maintenance and management text records is undoubtedly a time-consuming and laborious task. To quickly obtain the root cause text from unstructured data, this paper proposes an approach for root cause factor identification by using a root cause identification-new word sentence (RCI-NWS) keyword extraction method. The experimental results demonstrate that the extraction of railway fault text data can be performed using the keyword extraction method and the highest values are obtained using RCI-NWS.

Keywords: information extraction; root cause identification; railway fault; complex network; text data

1. Introduction

With the development of science and technology, the running speed of trains is rapidly increasing, which means that the capacity utilization of existing infrastructure is high [1]. Undoubtedly, railway safety also faces major challenges. High-speed trains may suffer from various faults during operation, such as human errors, internal equipment faults of the system and adverse external environments, which delays the trains. More seriously, such faults may cause accidents [2]. Through statistical research on train delays, it was found that the main causes of train faults are line faults, turnout faults, signal faults, car body faults, power supply faults and bad weather effects. Therefore, it is necessary to identify the root cause to help fault diagnosis, security monitoring and guidance for specific routine maintenance measures.

Much research has been proposed to identify the root cause, such as Fault Tree Analysis (FTA) [3], anticipatory failure determination (AFD) [4], subversion analysis [5], root conflicts analysis (RCA+) [6], etc. FTA is mainly used to construct the tree structure of accident faults and find out the cause of the accident failure. AFD is a method aimed at finding possible unexpected and undesirable events that risk disrupting the normal operation of a technical system, with the use of existing resources. Apart from AFD, subversion analysis is very useful in the identification of causes of given phenomena and events (faults, errors, negative impact), especially in situations where causes cannot be easily determined [5]. RCA+ is a technique for problem analysis developed for the top-down decomposition of problems to chains of causes and contradictions.

Recently, fault-related problems have attracted increasing attention from researchers in the field of railway transportation. Due to the difficulty in obtaining actual train operation data, studies at home and abroad mainly focus on the determination and prediction of faults, the simulation of faults

and the theoretical model of train delay propagation [7]. For example, as the railway track circuit is the most commonly used component for train detection worldwide, fault diagnosis for railway track circuits has already been addressed [8–11]. Moreover, because turnout is very important equipment in railway infrastructure, accounting for approximately 33% of the annual railway maintenance cost, several studies on turnout have been reported in the literature. A failure prediction algorithm of railway turnouts based on a neural network is proposed by Yilboga et al. [12]. Also, to predict railroad turnout failure, an autoregressive moving average model was developed [13]. Additionally, a simple state-based prediction method is proposed to detect and predict the fault progression of railway turnout systems [14].

In addition, various studies regarding methods for diagnosing railway system faults have been carried out in recent years. An artificial neural network (ANN)-based fault diagnosis method for a turning area of jointless track circuits was proposed by Zhao et al. [15]. Additionally, to detect faults in the railway track circuit, the neuro-fuzzy system and the Dempster–Shafer theory were employed [8]. Yin and Zhao proposed an automated vehicle onboard equipment (VOBE) diagnosis network for a high-speed train via a deep learning approach, which captures the complexity and uncertainty of the VOBE faults [16]. VOBE for trains can receive real-time data, including the speed limits, length of the track circuit, gradients and running speed, and then calculate the preferred control strategies [17]. Verbert et al. proposed such an approach for fault diagnosis in networks, which is knowledge-based and uses temporal, spatial, and spatiotemporal network dependencies as diagnostic features, requiring fault diagnosis methods that can work with a limited set of monitoring signals [18]. Bruin et al. proposed a long short-term memory (LSTM) recurrent neural network to identify faults based on commonly available measurement signals [19].

For the reason that most of the existing fault methods are based on state detection, starting from the mechanical properties of the equipment or the failure rate of the equipment, the root cause analysis performed by researchers is often based on structured data when it is necessary to consider unstructured text data, such as fault tracking reports, libraries, fault libraries, causal analysis, and process analysis. In addition to the identification of the causal factors required for risk analysis, past researchers have usually used manual processing or have neglected and then conducted related analysis on structured data. The identification of cause factors, once manual screening and identification are required, is undoubtedly a time-consuming and laborious task, especially when the data are large. Unstructured text data contain a considerable amount of useful information. The text data of railway faults, as an example, usually includes the root cause of the failure, the process, the main causal factors, the underlying factors, the relationship between the factors, and the consequences of the failure, and it requires considerable effort to identify the needed information. However, the content in unstructured data is not detailed.

Therefore, it is meaningful to study how to identify the cause of railway failures from the railway text data and to promote relevant subsequent analysis. Currently, there are many fault records and maintenance data in the maintenance and management of railway equipment. These unstructured data can play a very important role in equipment fault prediction. If we can make full use of these document data and excavate the underlying value of the data, it will greatly improve the reliability and security of railways.

Many researchers use word networks to research natural language processing. Beliga et al. showed that network-based methods have the best performance in extracting keywords [20]. Considering the similarity between the extraction of railway causal factors and the extraction of keywords, this paper considers whether the extraction of railway fault text data can be extracted by the keyword extraction method, thus effectively improving the extraction efficiency of accident causative factors. It provides a new idea for the analysis of the causes of railway accidents/events, the analysis of railway fault accidents and the risk analysis of railway faults. To reduce the time required and to use the text data correctly, which could be further used in research, we propose extracting the root cause of railway faults from every detail.

This paper is organized as follows. The principles that are employed in our proposed method are presented in Section 2. A method for identifying root causes is described in detail in Section 3. Section 4 shows the experimental analysis and relevant results. Finally, conclusions are summarized in Section 5.

2. The Base of the Proposed Method

2.1. Basic Statistical Properties of Complex Network

(1) Betweenness centrality

Betweenness centrality is defined to measure a node when it acts as a bridge in the shortest path between the other two nodes:

$$bc_i = \sum_{a \neq b \neq i} \frac{\phi_{ab}(i)}{\phi_{ab}} \tag{1}$$

where ϕ_{ab} is defined as the total number of shortest paths from node a to b, and $\phi_{ab}(i)$ is the number of shortest paths from node t to node k passing through node i. More details of these concepts are provided in [21,22].

(2) Closeness centrality

Closeness centrality is defined as the inverse of farness, which is the sum of the shortest distance between a node and all other nodes:

$$cc_i = \frac{N_{node} - 1}{\sum_{j-1}^{n} d(i,j)} \tag{2}$$

where N_{node} is the number of nodes in a network, $d(i,j)$ represents the distance between node i and node j, and n is the number of nodes in a network. More details of these concepts are provided in [21,22].

2.2. New Word Sentence (NWS) Keyword Extraction Method

The NWS method is a keyword extraction method proposed in our published paper [23], which is based on complex network theory. Figure 1 demonstrates that the NWS method includes two layers: a sentence network in the upper layer and the word network in the lower layer. In the sentence network, the square represents a node (sentence), while the circle represents a node (word) in the word network. The relationship between the same node is represented by a solid line. When a node (word) in a word network exists in a node (sentence) in the sentence network, their relationship can be built and represented by dotted lines. Therefore, we obtain the synthetic eigenvalue of each node (word) for the NWS method as follows:

$$E_i = \lambda WE_i + \gamma SWE_j \tag{3}$$

$$WE_i = \alpha \cdot bc_i + \beta \cdot cc_i \tag{4}$$

$$SWE_j = \omega_j \tag{5}$$

here E_i represents the eigenvalue of word i, WE_i represents the synthetic eigenvalue of word i, SWE_j represents the eigenvalue of sentence j, bc_i is betweenness centrality, cc_i is closeness centrality, ω_j is the node (word) contribution from node(sentence) contribution, $\lambda + \gamma = 1$ and $\alpha + \beta = 1$ [23].

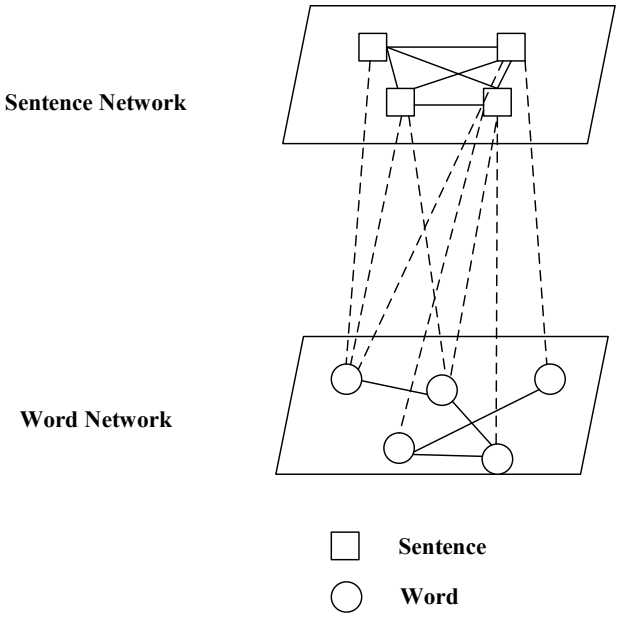


Figure 1. The structure of the new word sentence (NWS) method.

2.3. Apriori

An association rule is defined as the meaning of the form $X \Rightarrow Y$, where itemsets X is the front item and itemsets set Y is the right side. There are two conditions in an association rule: $X, Y \subseteq I$ and $X \cap Y = \emptyset$ supports determine how frequently the itemset appears in the dataset, while confidence determines how often an item has been found to be true. The definitions of support and confidence can be given in (6) and (7), respectively.

$$Support(\ X \Rightarrow Y) = \frac{\sigma(X, Y)}{N} \tag{6}$$

$$Confidence(X \Rightarrow Y) = \frac{\sigma(X, Y)}{\sigma(X)} \tag{7}$$

where σ is the summation notation and N represents the total number of all transactions [24,25].

2.4. Dependency Parsing

Dependency Parsing can be used to analyze the grammatical structure of sentences by establishing the relationship between different components in a sentence and identifying the grammatical structures of "subject verb" and "adverb". By analyzing the dependency relationship between language components, the syntactic structure of dependency analysis is revealed [26].

2.5. Evaluation Index

Precision (P):

$$P_n = \frac{1}{n}\sum_{i=1}^{n}\frac{|EC_i \cap MC_i|}{|EC_i|} \tag{8}$$

Recall (R):

$$R_n = \frac{1}{n}\sum_{i=1}^{n}\frac{|EC_i \cap MC_i|}{|MC_i|} \tag{9}$$

F-measure (F):

$$F_n = \frac{2P_nR_n}{P_n + R_n} \tag{10}$$

In Equations (8)–(10), n represents the number of fault records, EC_i represents the number of words automatically identified as a cause word by our method in record i, and MC_i represents the number of causes identified manually. $EC_i \cap MC_i$ represents the number of automatically identified cause factors that match the manually identified cause.

Similarity

Because of the differences between the algorithm in extracting information and manual operation, we take the similarity into consideration so that some situations (for example, driver's window) can be eliminated and used to calculate precision.

3. Proposed Method

To use the keyword extraction method to identify the causes, we present a method that employs the following assumptions:

Assumption 1: All fault records can be viewed as a fault report.

Assumption 2: The keyword extraction method can be applied to the text fault report and the assumptions of the method are also established.

Assumption 3: A failure cause is in the form of noun + verb combinations or independent nouns, where the latter noun could form a noun phrase with a *default*.

We reviewed the previous literature, and the cause of the failure text is in the form of noun + verb combinations or independent noun forms. Additionally, in a failure record, we found the cause is always the combination of a noun and a verb. Figure 2 shows a railway fault text record, which can have various parts of speech, such as nouns, verbs, and an uncertain number of them. In Figure 2, squares represent nouns, circles represent verbs, and other shapes represent the other parts of speech. When a noun (n.) and a verb (v.) can constitute a fault cause in this a record, such as n. + v., there will be shadows in both the square and circle, and they are marked as the "target" we are looking for.

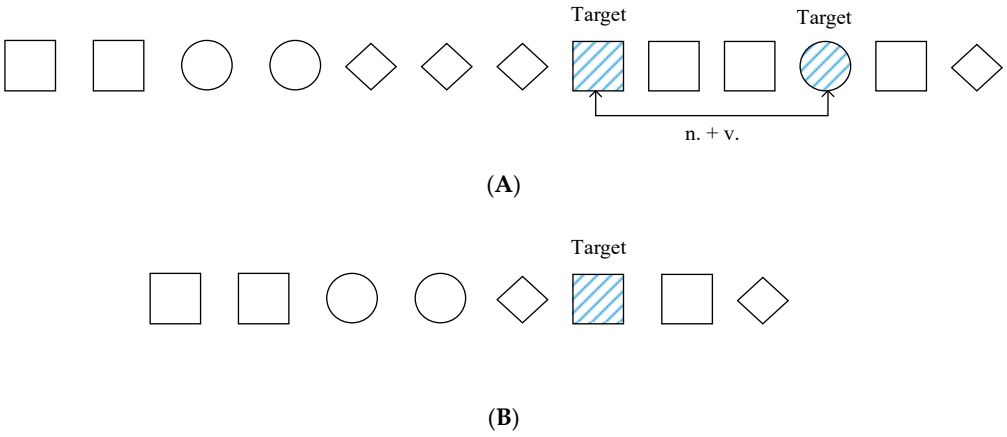


Figure 2. (**A**) A failure record with a combination target noun + verb (n. + v.). (**B**) A failure record with an insolate target (noun). Squares represent nouns, circles represent verbs, and other shapes represent the other parts of speech.

Therefore, we identified the result of the task to find the words in the form of n. + v. or isolate n.

To quickly find the cause noun and the cause verb, a dictionary is necessary. However, a dictionary of railway faults does not exist. To complete the task, we split the tasks of nouns and verbs.

Collecting and establishing a dictionary from previous literature is a good start but is not enough. Since the dictionary has incomplete content, it prevents us from identifying the cause nouns in the sentence (fault record). Figure 3 shows an incomplete dictionary with three hidden situations. Case 1: A useless word exists in the dictionary, which means this dictionary must filter the useless words. Case 2: A useful word should be detected but is not in the dictionary, which means the dictionary

is incomplete. Case 3: Both Case 1 and Case 2 occur. Basically, to solve these problems, we need to continuously expand the dictionary content through the task text and match the text to achieve the match function. Moreover, we also need to establish a text stopword list and filter and eliminate the text to supplement the dictionary. Incomplete words are reserved, and thus, filter functionality with the text is achieved. Therefore, the cause identification object is split into recognition reason nouns and verbs, and the identified tasks are split into establishing a cause dictionary and a deactivation dictionary.

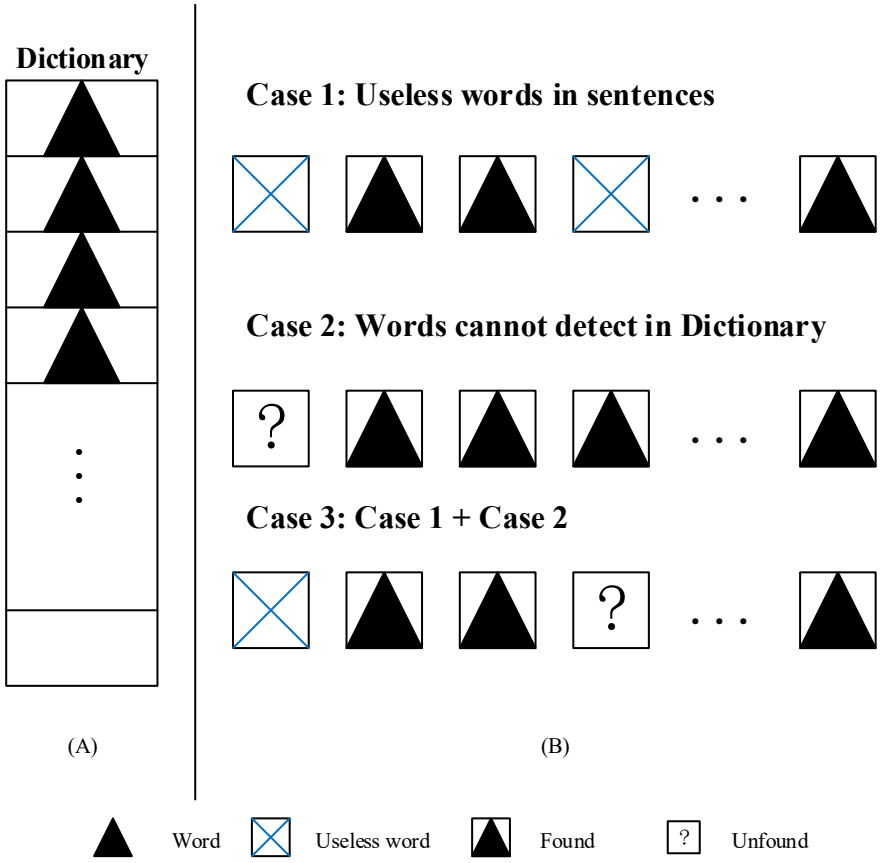


Figure 3. An incomplete dictionary with three hidden situations. (**A**) An incomplete dictionary. (**B**) Three hidden situations.

The steps of the proposed method are divided into the following: text preprocessing, noun recognition, and verb recognition. Noun recognition is further divided into establishing a user dictionary and stopword lists.

Briefly, extracting root cause text is divided into the following steps: identify the irrelated nouns; eliminate stopwords; require the suggested useful words; and clear the useful words to obtain the aim of our study. According to these steps, the root cause text should be extracted more precisely.

Specifically, the steps are as follows (see Figure 4):

Step 1 Pre-process the fault cause records as a fault cause text $FCR = \{cw_1, cw_2, cw_3, \ldots, cw_n\}$

Step 2 Text feature analysis and feature extraction.

Step 3 Follow the feature description and split the tasks into several processes. Collect the literature $DC = \{dc_1, dc_2, dc_3, \ldots, dc_i,\}$ and $LW = \left\{lw_1, lw_2, \ldots, lw_j\right\}$. Set up the first part dictionary $Dic_1 = \{DC, LW\}$.

Step 4 Build the second part dictionary $Dic_2 = \{kw_1, kw_2, kw_3, \ldots, kw_N\}$. Build a noun list, which selects the nouns from the cut text. Perform text keyword extraction on the fault cause text. Add the cause text from the top n keywords to the noun list.

Step 5 Build a stopword list $ST = \{ST_1, ST_2\}$. Consider the elemental information, such as train departure time and train station (location) and collect them as stopwords $ST_1 = \{Ti, Loc, Date, SN\}$. In addition, use the apriori algorithm to obtain the stopwords $ST_2 = \{st_1, st_2, st_3, \ldots, st_s\}$, which include frequency and meaningless words, to build a necessary stopword list.

Step 6 Merge the two dictionaries and obtain a new dictionary $Di = Dic_1 + Dic_2$, then filter this new dictionary by the stopword list $ST = \{ST_1, ST_2\}$ to obtain $ND = Dic_1 + Dic_2 - ST = \{d_1, d_2, d_3, \ldots, d_n\}$.

Step 7 Use the dictionary to match the cause noun. For word$_i$ in *FCR*, if word$_i$ is in ND, the cause word cw_i is obtained.

Step 8 Use a dependency parsing to find the cause verb cv_i if it exists; otherwise, the causal factor is combined with a *default* combination failure to obtain causal factor CF_i.

$$CF_i = \begin{cases} cw_i + cv_i\ , \text{ if } cv_i \text{ exists} \\ cw_i + default, \text{ if } cv_i \text{ does not exist} \end{cases}$$

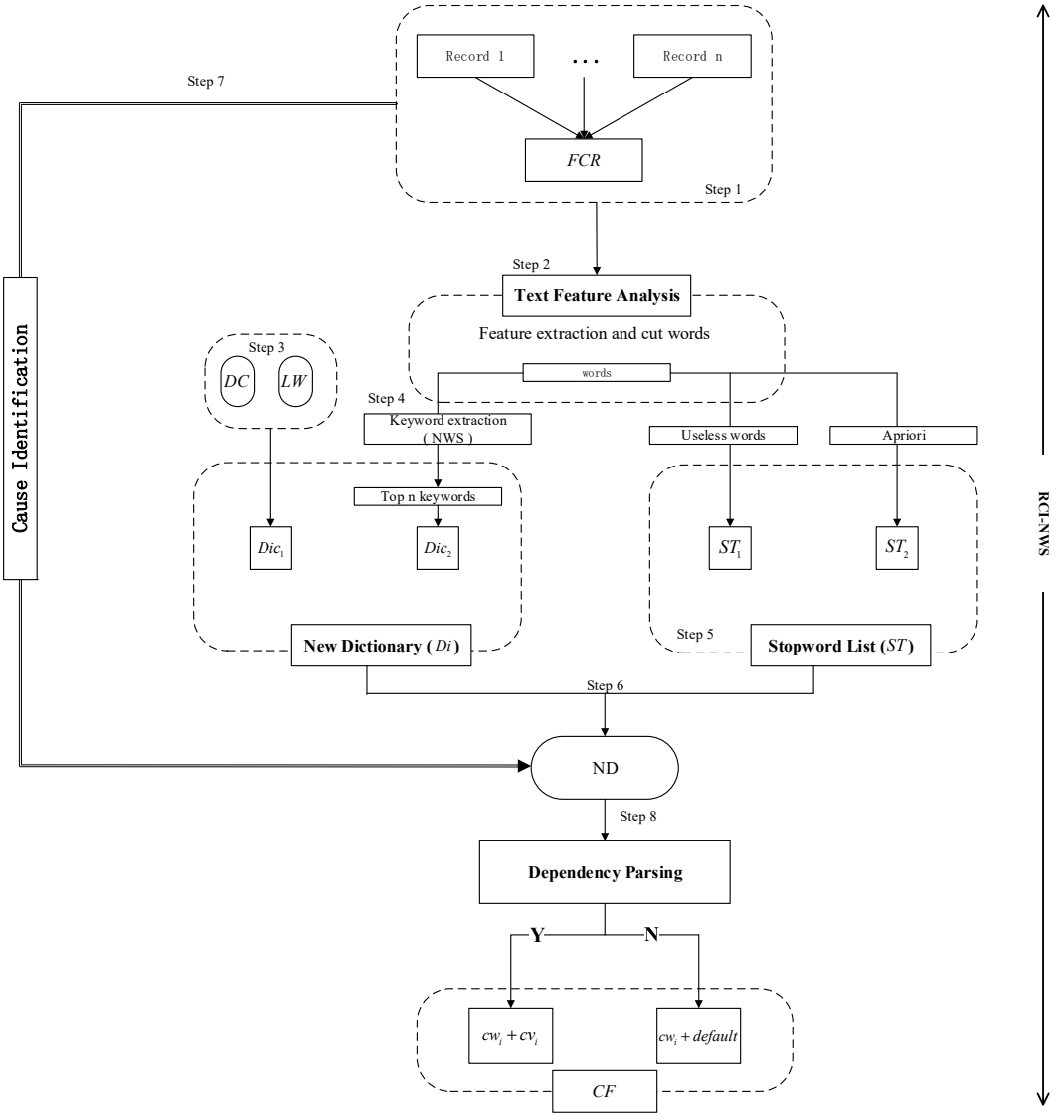


Figure 4. Framework of proposed method.

3.1. Feature Description and Selection

(1) Participle symbol feature

When the divided text is sparse, the resulting reason words are sparse; if the word segmentation is relatively close, the result may be composed of multiple nouns. Therefore, when the text is segmented, the characteristics of the word segmentation also affect the identification of the cause.

(2) Part of speech

By analyzing the fault causes manually extracted from previous literature, we find that these reasons are usually caused by a noun and a verb phrase or a single noun. The characteristic of these nouns is that they are all the structure of a train body or the name of a part of the train. Thus, the part-of-speech feature can be used to identify the root cause of railway faults. Thus, through the feature extraction of the text part of speech, we determined that the goal of the task is to find nouns and verbs related to the car body.

(3) Semantic feature

When analyzing the railway fault text data, summarizing the naming rules of the root cause identification is beneficial for improving the accuracy of the root cause. For example, "the traction current transformer charging resistor burns" can be composed of the body structure term "traction converter charging resistance" and the fault occurrence action "burning".

(4) Term length feature

Combined with Chinese writing habits, from the entire fault record, the description of the root cause of the fault will be in the second half of the sentence; that is, other information unrelated to the root cause identification is usually concentrated in the first half of the sentence. Thus, the word position of the reason word is a general and inconspicuous feature.

(5) Keyword characteristics

When we are extracting keywords from the text fault data, we see the keywords of "so", "discovery", "because cause", "have", "occur", "fault", etc. In conjunction with the sentence description of the fault in the text, many reason words occur after or before these feature words. These words are used to create a new dictionary of prompt words and to consider the use of the inverse maximum matching method for the location and semantic annotation of the reason words. These can be used as a feature in the base of text keyword features.

Therefore, to build a user dictionary, we consider the particularity of the railway fault text data, including the linguistic part of speech, semantics, keyword characteristics and other characteristics of the various levels of representation.

3.2. Identify the Root Cause Body Name

a Fundamental to the body list

We build the dictionary by combining the description of feature extraction in Section 3.1.

First, according to the characteristics of the reason words in the text, the reason can be divided into noun + verb form. For nouns, the nouns that constitute the reason are usually related to the train body structure. The fundamental method uses a dictionary to depict the desired words. However, most of the time, the records we have are limited and not open data, which means that we have to review the history and former literature for help with performing the extraction. Therefore, we obtained the preliminary version of Dic_1 of the body dictionary by finding limited literature.

Second, considering the keyword characteristics of the text, the word segmentation is obtained, and the nouns of the cause are extracted. The nouns after the word segmentation are extracted, and then the word length and sentence features are combined with the terminology of the sentence. According

to the characteristics of the keyword, the keyword must contain important reason words. The nouns related to the cause of the failure are found for Dic_2. Since the vehicle has been determined for the dictionary of body nouns, we consider the fault text for keyword extraction features. The keyword is extracted from the text describing the reason, and the keyword must contain important reason words. Combined with the feature analysis of the text, the dictionary additions are performed step by step, and all the dictionaries are merged to obtain the initial document *Di*.

b Stopword list foundation

Because the body list is still not sufficient for matching the target word in each record, to some extent a stopword list would enhance the efficiency of finding the root cause word. In a record, words such as city name, location, numbers and letters are eliminated immediately. To find more valid stopwords, we use the keyword extraction method and then select stopwords from the top 100 keyword list. In addition, we also use apriori to collect more stopwords and update the original stopwords list.

c Merge

The obtained dictionary file and the stopwords document are processed by the merchant, thereby obtaining a dictionary in which we can identify the reason noun.

3.3. Identify the Root Cause Action

When we obtain the reason noun, we start the second stage of identifying the root cause action. We introduce a dependency parsing, find pairs of grammatical verbs through existing nouns, and perform similarity calculations on these verbs with the obtained verb lists to filter out effective verbs and finally obtain the reason, a phrase of nouns and verbs.

4. Experimental Results

In this section, to verify the validity of the proposed method, we manually extracted the reason for the 299 fault texts, which were compared with the fault reason extracted using our method. Moreover, we adopted traditional indicators, including the precision, recall, and F-measure, to evaluate the performance of the algorithms.

4.1. Fault Text Data

The data used were from a seven-month railway record and were collected in China from January to July, 2010. The contents of the database include the dates, the railway system information, the detailed description of the faults, the miles, and the consequences of the faults. The detailed description of the faults contains 299 descriptive text records of faults. Each descriptive record provides almost all the information related to a fault in detail, such as railway name, time and causal factors in a descriptive sentence.

As a short passage consisting of 299 Chinese railway records, the basic process for keyword extraction contains parts of speech, ranking the values and finally, obtaining the required keywords. In this short passage, we selected the top 20 keywords in the experiment while we could easily have the mix text words, which means some useful words and useless words among the top 20 keywords. (see Table 1).

Table 1. Top 20 keywords in NWS methods (translated to English).

Keyword
attachment
Wuhan Bureau
Emu
fault
run
warehouse
maintenance
Jinan Bureau
check
Shanghai
configuration
way
Beijing Bureau
Foreign
Shenyang Bureau
Traction Converters
Nanchang
code
driver's room
burn loss

4.2. Cause Factor Analysis: Characteristics

We randomly selected certain records (sentences) to extract root cause factors, and some common characteristics were as follows:

(1) A root cause factor could be divided into two types: one is a noun with a verb, and the other is a noun with a "fault" in Chinese. To better describe and distinguish, we call the noun the root cause word.
(2) The part of speech of the root cause word is a noun, and the noun is related to the train structure.
(3) The average length of the root cause word is approximately two; in other words, its length is greater than one word.
(4) A root cause word is also important when all records are regarded as similar to one in a passage. Therefore, when extracting keywords in such passages, root cause words are more likely in the keyword list.
(5) Ninety percent of the root cause words are in the 37% length of a sentence.

In this experiment, we had two types of study references. The first is the train technical terms (TTTs) as a dictionary with more than 6188 terms translated into Chinese as well as English. As a dictionary, it has many Chinese terms, where the length of the Chinese terms ranges from two (door lock) to nine (cyclone-type air filter) in TTTs and contains almost all of the train body. Second, with the help of these studies, we can directly or indirectly apply these causal factors for analysis, which are summarized and divided into five classes: track, roadbed and structures; signal and communication; mechanical and electrical faults; human; and other. These five classes contain 326 factors recorded by the structure with noun and verb in a record, which were generated by workers, machinery, electrical equipment, external environment, etc.

4.3. Identify the Root Cause

To illustrate the role of the keyword extraction method in building a dictionary, we designed Experiment 1 and Experiment 2, and the difference was whether the keyword extraction method was considered. Experiment 2 considered the cause recognition results of different keyword extraction

methods. Experiment 3 considered the effect of stopping the vocabulary on the cause identification. Experiment 4 considered the effect of text similarity on the cause of fault cause recognition (see Figure 4).

To illustrate the role of the keyword extraction method in the proposed method, we used the most frequent (MF) method to take the place of NWS in the part of root cause identification-new word sentence (RCI-NWS).

4.4. The Most Frequent (MF) method

In the MF keyword extraction method [27,28], the score of the word w_i is obtained by counting the number of occurrences of the word in the matrix:

$$score(w_i) = freq(w_i) = \sum_{s_k \in S} occurence_{w_i(s_k)} \tag{11}$$

where w_i represents word i, s_k represents a sentence k, S denotes a total sentences set and $w_i(s_k)$ denotes word i in sentence k.

In addition, we compare the results considering stopwords and word similarity.

The letters R, N, and M indicate the keyword extraction method used in the dictionary section. R is the abbreviation for raw, which means that the dictionary is not built using any keyword extraction method, and only the first-generation dictionary newly created by the literature is used. N is the NWS keyword extraction method, and M is the MF keyword method. The number after the letter indicates whether the factor is considered in the method, whether one considers the stopword table for the merge operation, and whether the other considers the text similarity. For example, R00 indicates that the fault text reason identification uses only the first-generation dictionary and does not consider the stopword list and similarity; N11 indicates that the NWS keyword extraction method is used in this text failure cause identification to create a new dictionary and the stopword is used. The vocabulary is used to fuse the dictionary, and the text similarity is considered in the experimental calculation.

To illustrate the role of keyword extraction in the new dictionary, we compared Step 2 using only the first-generation dictionary and using NWS as the keyword extraction. The comparison results are shown in Figure 5 (R00, N00), and it can be seen that NWS is used when the keyword is extracted because its extraction is better than using only the first-generation dictionary. To illustrate the role of the stopword list in the new dictionary, we performed two sets of comparisons ((R10,R00), (N10,N00)), and the results showed whether the keyword extraction NWS method was used during the dictionary process. All of them show that considering the stopword list, the effect of extracting the cause of the fault is better than not considering the stopword list. However, the N10 effect is far superior to R10. This comparison shows that the method of keyword extraction is helpful in identifying the cause words (R10, N10) in the fault text.

To further illustrate the performance of the NWS keyword extraction method in fault cause text recognition, we chose the MF keyword method. The performance of MF in the literature is more prominent. When we compare the performance of NWS and MF, we also carried out three sets of comparison implementations: (1) M00 and N00; (2) M10 and N10; and (3) M11 and N11. As can be observed from Figure 5, the results of our proposed method are better than those of MF, regardless of whether stopwords and similarities are considered.

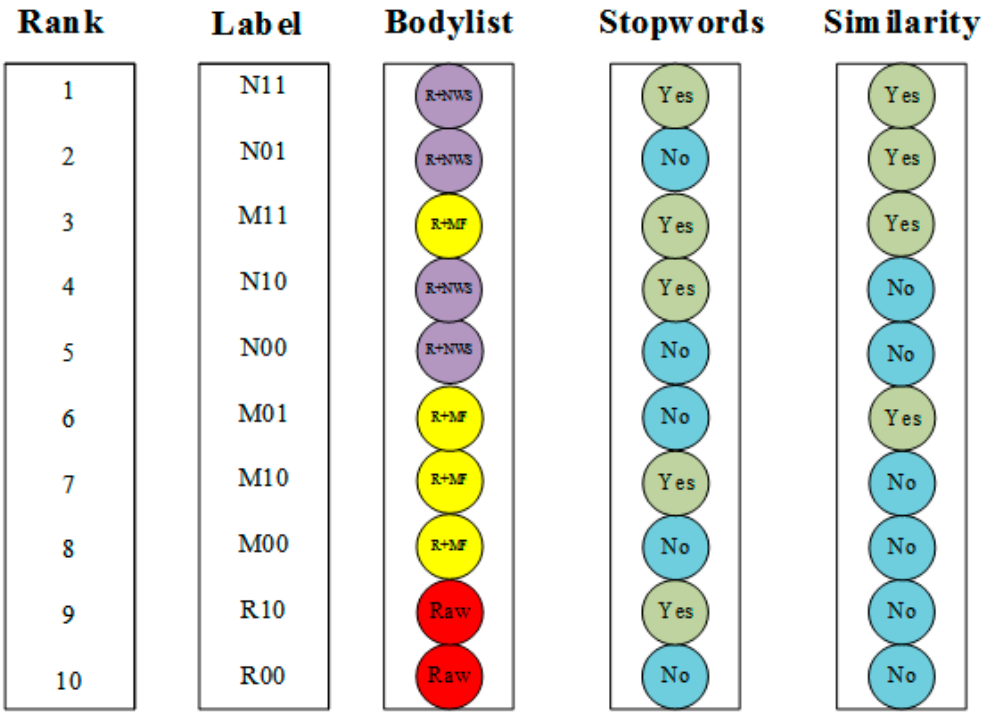


Figure 5. The difference between each experiment. MF = most frequent; M = MF method; R = raw N = NWS method.

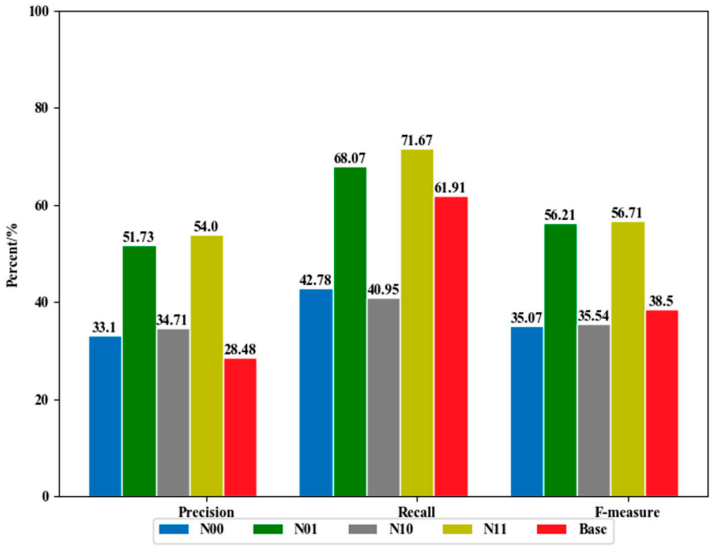
To investigate the influence of root cause noun extraction on the performance, we conducted experiments with different methods, stopword lists and similarities. Moreover, we added the base to compare with others. The base is the average level based on the manual performance of root cause noun extraction. It is defined as a sample with the possibility of good performance and is introduced to reflect the better performance if the values of P, R, and F are higher than the base values.

Table 2 shows that the performance of N11 is the best, which represents the performance of our proposed method. The highest values in RCI-NWS are M11, which means that in this situation, the body list was required by the RCI-NWS method, and both considered the two changing differences in stoplist and similarity. Specifically, comparing M11 and N11, the gap in precision increases to 8.72%, the recall increases from 47.28% to 54% and from 37.72% to 71.67%, and the F-measure values increase from 39.68% to 56.71%. All three improvements in N11 are higher than the average performance corresponding to the base.

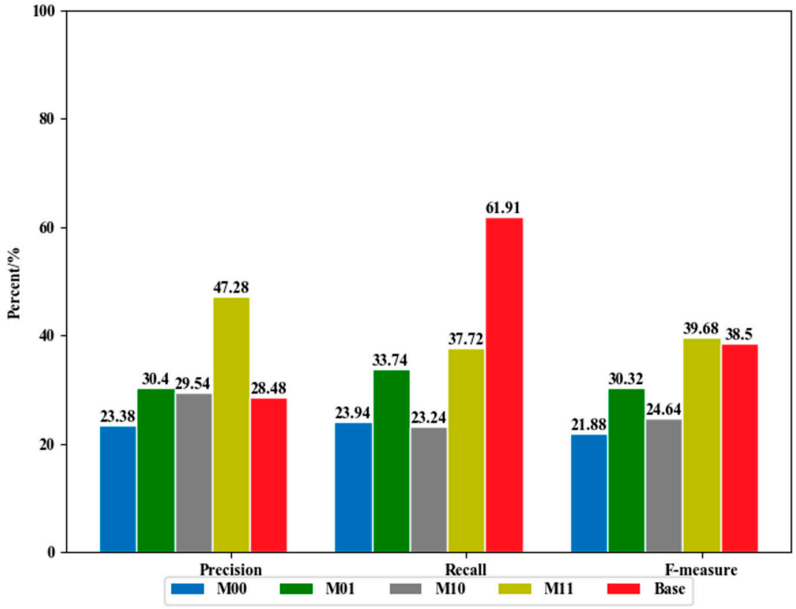
Table 2. Results in various conditions. P = precision; R = recall; F = F-measure.

Name	P	R	F
M00	23.38%	23.94%	21.88%
M01	30.40%	33.74%	30.32%
M10	29.54%	23.24%	24.64%
M11	47.28%	37.72%	39.68%
N00	33.10%	42.78%	35.07%
N01	51.73%	68.07%	56.21%
N10	34.71%	40.95%	35.54%
N11	**54.00%**	**71.67%**	**56.71%**
R00	20.22%	15.09%	16.08%
R10	20.94%	14.60%	16.21%
Base	28.48%	61.91%	38.50%

To better understand the relative performances of the RCI-NWS and RCI-MF methods, the trends in precision, recall, and F-measure values are shown in Figure 6. From Figure 6A, we can see that the bars of N01 (green) and N11 (yellow) are both always higher than the bars of base (red), which represents RCI-NWS using the NWS keyword extraction method. If the NWS keyword extraction method combines with a stoplist and similarity, the performance of N11 is the best. Similar results are shown in Figure 6B, but the performance of M11 demonstrates that its recall performance is not higher than that of the base. Furthermore, it can be observed from Figure 6C, comparing the best performance in RCI-NWS (M11), RCI-MF (N11), base and original (R00,R01), the highest value of the three is for M11 (red); therefore, our proposed method exhibits the best performance.

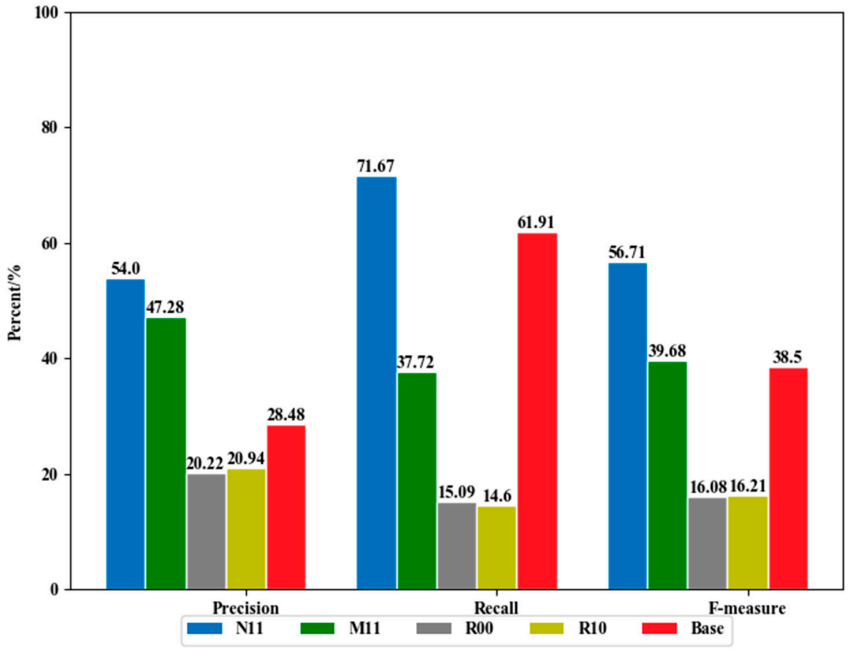


(**A**)



(**B**)

Figure 6. *Cont.*



(**C**)

Figure 6. Performance of the root cause identification (RCI)-NWS and RCI-MF methods. (**A**) Comparative performance of RCI-NWS methods in different conditions; (**B**) Comparative performance of RCI-MF methods in different conditions; (**C**) Comparative performance among best performances in RCI-NWS, best performances in RCI-MF, Raw and Base.

5. Conclusions

In this paper, a novel approach for extracting root cause factors was proposed. For this purpose, a detailed procedure was provided to build an effective dictionary by utilizing combination of the keyword extraction method NWS, Apriori and Dependency Parser. To demonstrate the application of the proposed approach in practice, Chinese railway fault records were selected and investigated. By employing the proposed approach, the results demonstrate that applying our approach successfully achieves the goal of extracting the root cause factors, which can further influence future analysis and clustering among these results, and reducing the time required to use the texts correctly, which could be further used in research. The extraction of railway fault text data can be extracted by the keyword extraction method, which provides a new idea for the analysis of the causes of railway accidents/events, the analysis of railway fault accidents and the risk analysis of railway faults.

However, our proposed method can be further improved. The method considers only the root cause identification from railway faults, which are related to the railway file; thus, future work can make full use of different traffic-related fields. In addition, in future research, we intend to apply our proposed method to conduct research on cause chain identification.

Author Contributions: Conceptualization, L.Y. and K.L.; methodology, L.Y. and D.Z.; software, L.Y. and D.Z.; writing—original draft preparation, L.Y.; writing—review and editing, K.L., L.Y., S.G. and D.Y.; supervision, K.L.

Acknowledgments: This work was supported by the National Natural Science Foundation of China (Grant No. 71621001), the National Key Research and Development Program of China (Grant No. 2017YFB1201105), the Research Foundation of State Key Laboratory of Railway Traffic Control and Safety (Grant No. RCS2019ZT001), and the Fundamental Research Funds for the Central Universities (2018YJS204).

Conflicts of Interest: No potential conflict of interest was reported by the authors.

Nomenclature

bc_i	betweenness centrality
ϕ_{ab}	total number of shortest paths from node a to b
$\phi_{ab}(i)$	the number of shortest paths from node a to node b through node i
cc_i	closeness centrality
N_{node}	the number of nodes in a network
$d(i, j)$	the distance between node i and node j
E_i	the comprehensive eigenvalue of word i
WE_i	the eigenvalue of word i
SWE_j	the eigenvalue of sentence j
λ	coefficient
γ	coefficient
α	coefficient
β	coefficient
I	a set of n binary attributes called items
X	antecedent (left-hand-side or LHS)
Y	consequent (right-hand-side or RHS)
σ	summation notation
$Support(X \Rightarrow Y)$	how often a rule is applicable to a given dataset
$Confidence(X \Rightarrow Y)$	how frequently items in Y appear in transactions that contain X.
P_n	Precision
R_n	Recall
F_n	F-measure
W_i	word i
EC_i	the number of words automatically identified as cause word by our method in record i
MC_i	the number of causes identified manually.
cv_i	cause verb
DC	Dictionary
Base	the average level based on the manual performance of root cause noun extraction.
Raw	the dictionary is not built using any keyword extraction method
Score	count the number of occurrences of the word in the matrix
Occurrence	the number of occurrences of a word
Default	a fixed word combined with an isolate noun
Dic_1	the first part dictionary
Dic_2	the second part dictionary
ST	stopword list
ST_1	stopword list 1
ST_2	stopword list 2
D_i	new dictionary
NWD	new word Dictionary
NWS	new word sentence keyword extraction method
RCI-NWS	root cause identification-new word sentence method

References

1. Yang, X.; Li, X.; Ning, B.; Tang, T. A survey on energy-efficient train operation for urban rail transit. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 2–13. [CrossRef]
2. Wen, C.; Li, Z.; Lessan, J.; Fu, L.; Jiang, C. Statistical investigation on train primary delay based on real records: Evidence from Wuhan–Guangzhou HSR. *Int. J. Rail Transp.* **2017**, *5*, 170–189. [CrossRef]
3. Lin, C.Y.; Saat, M.R.; Barkan, C.P. Fault Tree Analysis of Adjacent Track Accidents on Shared-Use Rail Corridors. *Transp. Res. Rec. J. Transp. Res. Board* **2016**, *2546*, 129–136. [CrossRef]

4. Chybowski, L.; Gawdzińska, K.; Souchkov, V. Applying the Anticipatory Failure Determination at a Very Early Stage of a System'S Development: Overview and Case Study. *Multidiscip. Asp. Prod. Eng.* **2018**, *1*, 205–215. [CrossRef]
5. Chybowski, L.; Bejger, A.; Gawdzińska, K. Application of Subversion Analysis in the Search for the Causes of Cracking in a Marine Engine Injector Nozzle. *World Acad. Sci. Eng. Technol. Int. J. Ind. Manuf. Eng.* **2018**, *12*, 302–308.
6. Souchkov, V. Application of Root Conflict Analysis (RCA+) to formulate inventive problems in the maritime industry. *Sci. J. Marit. Univ. Szczec. Zesz. Nauk. Akad. Morska W Szczec.* **2017**, *51*, 9–17. [CrossRef]
7. Ping, H.; Qiyuan, P.; Chao, W.; hongcan, L.I.Z. Study on high-speed railway disruption classification and model of its influence on train number. *China Saf. Sci. J.* **2018**, *28*, 46–53.
8. Oukhellou, L.; Debiolles, A.; Denoeux, T.; Aknin, P. Fault diagnosis in railway track circuits using Dempster–Shafer classifier fusion. *Eng. Appl. Artif. Intell.* **2010**, *23*, 117–128. [CrossRef]
9. Cherfi, Z.; Oukhellou, L.; Côme, E.; Denoeux, T.; Aknin, P. Partially supervised independent factor analysis using soft labels elicited from multiple experts: application to railway track circuit diagnosis. *Soft Comput.* **2012**, *16*, 741–754. [CrossRef]
10. Sandidzadeh, M.; Dehghani, M. Intelligent condition monitoring of railway signaling in train detection subsystems. *J. Intell. Fuzzy Syst.* **2013**, *24*, 859–869.
11. Lin-Hai, Z.; Jian-Ping, W.; Yi-Kui, R. Fault diagnosis for track circuit using AOKTFRs and AGA. *Control Eng. Pract.* **2012**, *20*, 1270–1280. [CrossRef]
12. Yilboga, H.; Eker, O.F.; Guclu, A.; Camci, F. Failure prediction on railway turnouts using time delay neural networks. In Proceedings of the 2010 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications (CIMSA), Taranto, Italy, 6–8 September 2010; pp. 134–137.
13. Guclu, A.; Yilboga, H.; Eker, Ö.F.; Camci, F.; Jennions, I.K. Prognostics with Autoregressive Moving Average for Railway Turnouts. In Proceedings of the Annual Conference of the Prognostics and Health Management Society, PHM, Macau, China, 12–14 January 2010.
14. Eker, O.F.; Camci, F. State-based prognostics with state duration information. *Qual. Reliab. Eng. Int.* **2013**, *29*, 465–476. [CrossRef]
15. Zhao, L.; Zhang, C.; Qiu, K.; Ling, Q. A fault diagnosis method for the tuning area of jointless track circuits based on a neural network. *Proc. Inst. Mech. Eng. F J. Rail Rapid Transit.* **2013**, *227*, 333–343. [CrossRef]
16. Yin, J.; Zhao, W. Fault diagnosis network design for vehicle on-board equipments of high-speed railway: A deep learning approach. *Eng. Appl. Artif. Intell.* **2016**, *56*, 250–259. [CrossRef]
17. Yin, J.; Chen, D.; Li, L. Intelligent train operation algorithms for subway by expert system and reinforcement learning. *IEEE Trans. Intell. Transp.* **2014**, *15*, 2561–2571. [CrossRef]
18. Verbert, K.; De Schutter, B.; Babuška, R. Exploiting spatial and temporal dependencies to enhance fault diagnosis: Application to railway track circuits. In Proceedings of the 14th European Control Conference, Linz, Austria, 15–17 July 2015; pp. 3052–3057.
19. De Bruin, T.; Verbert, K.; Babuska, R. Railway track circuit fault diagnosis using recurrent neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2017**, *28*, 523–533. [CrossRef] [PubMed]
20. Beliga, S.; Martinčić-Ipšić, S. Node selectivity as a measure for graph-based keyword extraction in Croatian news. *Vet. Microbiol.* **2014**, *152*, 235–246.
21. Chen, Q.; Jiang, Z.; Bian, J. Chinese keyword extraction using semantically weighted Sixth international conference on intelligent human–machine systems and cybernetics. *IEEE Comput. Soc.* **2014**, *2*, 83–86.
22. Beliga, S.; Meštrović, A.; Martinčić-Ipšić, S. An overview of graph-based keyword extraction methods and approaches. *J. Inf. Organ. Sci.* **2015**, *39*, 1–20.
23. Yang, L.; Li, K.; Huang, H. A new network model for extracting text keywords. *Scientometrics* **2018**, *116*, 339–361. [CrossRef]
24. Doostan, M.; Chowdhury, B.H. Power distribution system fault cause analysis by using association rule mining. *Electr. Power Syst. Res.* **2017**, *152*, 140–147. [CrossRef]
25. Tan, P.N.; Steinbach, M.; Kumar, V. Association analysis: Basic concepts and algorithms. In *Introduction to Data Mining*; Addison-Wesley: Boston, MA, USA, 2015; pp. 327–414.
26. Nugues, P.M. Dependency Parsing. In *Language Processing with Perl and Prolog. Cognitive Technologies*; Springer: Berlin/Heidelberg, Germany, 2014.

27. Rossi, R.G.; Marcacini, R.M.; Rezende, S.O. Analysis of domain independent statisticalkeyword extraction methods for incremental clustering. *Learn. Nonlinear Models* **2014**, *12*, 1737. [CrossRef]
28. Onan, A.; Korukoğlu, V.; Bulut, H. Ensemble of keyword extraction methods and classifiers in text classification. *Expert Syst. Appl.* **2016**, *15*, 232–247. [CrossRef]