Defect Text Mining Technique and Application in Power Grid Based on the modified Semantic Framework

T M Yang1,2, Y H Wang1, Z T Han1, Y Liang1, J Gao1 and X Ji1

1Economy Technical Research Institute of Liaoning Electric Co. Ltd., Shenyang 110015, China

E-mail: 18240369315@163.com

Abstract. There are mass data that contain important defect texts in the power grid enterprise, and they contain important reliability information. And the efficiency is very low to mine the exact information about the texts especially when the texts are in Chinese. Thus, the defect text mining technique based on the modified semantic framework is proposed. All texts are translated into English and use the text mining model based on the modified semantic framework, the defect texts are divided into a fixed pattern and the digital information can be extracted accurately. Take the transformer as an example, the first step is to establish the ontology dictionary and to separate the sentence and extract the texts’ features. Then, the modified power semantic framework and the semantic slots are defined, and the slots filling method and the semantic framework construction process are discussed, which can automatically perfect the ontology dictionary by merging the word series. Finally, the researches of defect text mining results of statistical reliability are studied, and the results show that the proposed model and method is feasible and effective when applied to automatic classification and statistics of grid defect.

1. Introduction

Big data technology is the research hotspots in many professional fields. Big data includes the structured data and unstructured data. The unstructured data contains texts, audio files, images and so on. And the information mining of unstructured data is very difficult. Since many defect information of power equipments are recorded in texts, it always take a long time to analyze all the defect texts and it is extreme difficult to mine effective information. To classify the defects and analyze all reliability information need to be conducted manually, which consume time and energy. Therefore, the texts mining technique is very important and urgent.

There are some researches on text mining technique in different professional fields. And it is difficult because the text mining technique should combine with related professional knowledge [1]. In the mechanical field, some scholars use the natural processing technique to analyze the texts based on plenty of history diagnostic reports. And through comparing with the describing texts of equipments’ states, the most similar situation is found and provides the diagnostic recommendations [2]. In electrical field, there are scholars proposed information mining technology by using machine learning algorithm to mine quantities of defect data, which can provide basis for faults forecasting and preventive requirement of power equipments [3]. The describing texts in power system usually originate from manual writing, which have a difficult syntactic structure and is quite hard to distinguish the subject-predicate-object component [4-5].

In other electrical period, the text mining technique are used in the automatic generation of work
order and operation order, and the semantic framework is proposed to improve the generation
efficiency of work order and operation order [6]. And in [7], the dynamic template scheme is proposed
to evaluate the safety assessment, which can improve the data mining efficiency in work order
and operation order. In the existing research, the vector space method is proposed for mining information.
However, it can’t extract digital information and quantifier and can only classify the grade of defects.

Therefore, in order to realize deeper information mining, the texts mining technique is proposed
based on the modified semantic framework and the power defect texts are chosen as researching
objects. And based on the modified semantic framework model, the exact digital information and
quantifier can be extracted. The remainder of this paper is organized as follows. The ontology
dictionary of faults and defects is established and the power semantic framework and slots are defined
in Section II. The process of semantic framework construction and slots filling is proposed in Section
III. Section IV carries on a case study and make the reliability and defects statistics based on texts
mining. Finally, the final conclusions are provided in Section V.

2. The defect text mining technique based on the modified semantic framework

2.1. The definition of power semantic framework and semantic slots

After getting the defect texts from the internet or the information management system of the power
grid, the defect texts should be analyzed. And based on the analysis of large amounts of defect texts
[8-9], we can find that only one defect text can describe various defects. If the defects are not
separated, there will exist confusion of understanding and reduce the statistical accuracy. Therefore,
the semantic framework and semantic slots are proposed in [8] to solve these confusions of defect
texts in Chinese. Since not all defect texts are in Chinese and there exist various interpretations of
Chinese sentence, a modified semantic framework is proposed in this paper. All defect texts are
translated into English firstly. In the field of frame semantics, the sentence frame is the background
and reason of word understanding and it can present the whole meaning of one text. In the expert
system, the framework is a data structure of formal description, and it is composing of several
semantic slots with strong relevance. The semantic slots are composed of two parts, which are core
elements and non-core elements. Therefore, the semantic framework can be regarded as the minimum
unit to describe the defects.

Usually, a complete semantic framework is composed of four slots, which can be described as
follow,

\[ F = \{ C_L, C_S, A, D \} \]  \hspace{1cm} (1)

Where, \( C_L \) represents the large component and \( C_S \) represents the small component, \( A \) represents
the attribute of the components and \( D \) represents the degree of defects. And in the semantic framework, \( C_S \)
and \( A \) are core elements while \( C_L \) and \( D \) are non-core elements.

For example, one defect text is translated or described as “The radiator of No.2 main transformer
cooler has oil leakage one drop per minute.” Based on the modified semantic framework, this text can
be separated into four parts to fill in the semantic slots as is shown in Table 1.

| Table 1. An example of the semantic slots in semantic framework |
|---------------------------------------------------------------|
|                  | \( C_L \) | \( C_S \) | \( A \)             | \( D \)               |
| No.2 main transformer cooler | The radiator | Oil leakage | One drop per minute |

Therefore, the defect texts can be separated into four parts according to its part of speech. And
according to the type of semantic slots, the part of speech can be described as \( C_L \), \( C_S \), \( A \), \( D \), which can
be regarded as the basis of slots-filling method. If one defect text contains several defects, it will
produce more than one semantic framework.

2.2. The mining model of power defect texts based on the semantic framework
In order to mine the important information in defect texts, a mining model of Chinese text aiming at statistical application in power grid is proposed in [1]. Since the model is only applied in Chinese texts, a modified model is proposed to be able to be applied in both English and Chinese texts mining technique. The model contains three layers, which are pre-processing layer, processing layer and statistical application layer.

a) The pre-processing layer
The pre-processing layer is to analyze all history fault records, which is aiming at obtaining the structure of defect text and establishing the ontology dictionary. This is the basic work of text mining, and it contains the follow contents.

1) Terms separating technique
There are several terms separating techniques, and in this paper the hidden Markov model is chosen.
2) The extraction of terms’ attributes
It contains three steps: Firstly, to count the frequency of each term in the defect text and sort them by frequency. And some special terms like the symbols, human names and place names are eliminated. The second step is to extract the common attributes. The possibility of the four slots to exist at the same time is very low, and the non-core elements are always missing. The third step is to extract the position of terms, and there exists strong regularity of the position of the four slots, and the typical order is \{C_L, C_S, A, D\} or \{C_L, C_S, D, A\}.
3) Establish the ontology dictionary
Based on the national standard, guideline and relative rules, we can sort all high frequency terms and make classification and establish the ontology dictionary. Since the ontology dictionary should be established manually, it will require a huge workload. What’s more, the richness of the ontology dictionary will directly influence the accuracy of mining the defect texts. There is an example of the ontology dictionary in Table II. And set the ontology term as the center, the ontology dictionary also contains part of speech, homonym, and synonym and so on. The basic framework of one word can be described as \{ontology term, part of speech, instance 1, instance 2,…\}.

| Ontology term | Part of speech | Instance 1                  | Instance 2                  |
|---------------|----------------|-----------------------------|-----------------------------|
| Oil filter    | C_S            | Oil filtering device        | On-line oil filter          |
| Relay         | C_S            | Oil pressure relay          | Phase sequence protection relay |

b) The processing layer
The processing layer is to finish the semantic slots filling, the construction of semantic framework and the combination of words of one concrete defect text.

1) Slots filling. Based on the ontology dictionary, the words with actual meanings are extracted and filled in the slots. And the words with actual meanings can be divided into two types, which are texts and numerical symbols [10].
2) The construction of semantic framework. Based on the close-range matching method, the words and each semantic slot match according to the semantic relevance degree. Since the semantic framework is composed of four slots and if there exist the condition of slots missing, there will be 2 to 4 slots.
3) The combination of words. In a fixed pattern, several continuous slots C_S can be combined as one slot, which can be described as \(C_{S_1}C_{S_2}C_{S_3}…C_{S_n}=C_S\). Meanwhile, \(C_S\) can be conducted as the instance of \(C_{S_n}\) in the ontology thesaurus.

c) The statistics and application layer
The unstructured data turned out to be structured data after the pre-processing and processing layer, and every defect text is turned out to be a fixed pattern, which can be understood by those search engines or softwares. In this paper, it is applied in the defects statistics.

The statistics and application layer is to finish the reliability and defects statistics. For example, we
can count the defect types and amount of specific equipment. What’s more, we can count the defect types and amount of one specific type equipment in an area or in a specific time period. These statistics results can provide reliability evaluation basis for equipment design, production and operation.

3. **The process of semantic framework construction and slots filling**

3.1. **The slots filling**

The slots filling process contains two parts, which are texts searching and regular expression. The texts searching method set the word string length in the dictionary as the normal length. From the beginning of the first word, take the normal length of word string as a comparison to match with the strings in the dictionary. If the strings are matched, the slot can be filled with the part of speech. If not matched, then move down the position and compare with the next string. According to the relation between the ontology term and its instance, and the homonym and synonym can be distinguished while slots filling. The regular expression is used to extract numerical words which match the regular expression grammar, and to get the standard expression of numerical words.

The flow chart of slots filling process is shown in Figure 1. Where, the position of word is the sequence of the word in the sentence. \( N \) represents the total number of words.

The initial semantic framework is \( F = \{C_L, C_S, A, D\} \), and \( C_L \) can be described as \( C_L = \{w_1, w_2, \ldots, w_m, \ldots, w_M\}^T \), where \( w_m \) contains the information of words and position, \( M \) represents the number of part of speech. And \( C_S, A \) and \( D \) has the same structure with \( C_L \).

![Flow chart of slots filling](image)

**Figure 1.** Flow chart of slots filling

3.2. **The construction of semantic framework**

Firstly, take the core element \( E_1 \) as the basic point, and the number of words between the non-core
element $E_2$ and $E_1$ is defined as $\text{Dis} \left( E_1, E_2 \right)$, the smaller the $\text{Dis} \left( E_1, E_2 \right)$ is, and the relationship between $E_1$ and $E_2$ is stronger. Therefore, the nearest $E_1$ is chosen to match with $E_2$.

Based on the rules and the typical order between the flots ($C_L, C_S, A, D$ or $C_L, C_S, D, A$) above, a specific semantic framework construction method is discussed. Firstly, if there exist the core element $C_S$, then towards the left to match $C_L$ and towards right to match $A$ and $D$. If there doesn’t exist $C_S$, and then set $A$ as the center and towards left to match $C_L$ and $D$. The flow chart is shown in Figure 2.

![Flow chart of constructing semantic framework](image)

Figure 2. Flow chart of constructing semantic framework

Where, the goal to judge $C_S$ empty or not is to set the core element; $K$ represents the number of core element column.

The middle semantic framework turned out to be the final framework through the strings combination. And the combined strings can be used to expand the ontology dictionary. The process of strings combination is as follow, from the start to compare the adjacent line of the middle framework. If the slots of $C_L$, $A$, $D$ are all the same and these two lines can be combined.

4. Case Study
The case chooses around 3 thousand texts of transformers’ defects collected through web crawler in 2012-2017 and all texts are translated into English. Based on the pre-process of these defect texts, the ontology dictionary is established. In the ontology dictionary, there are 389 ontology terms and the number of $C_L$, $C_S$, $A$, $D$ is 5,108, 231 and 45. Firstly, the mining process of one typical defect text is described below. Secondly, the statistical analysis is conducted after all statistic results.

4.1. The mining process of one specific defect text
One specific defect text is described as “The respirator silica gel of No.1 main transformer changed color, its radiator has oil leakage one drop per minute, and the on-load tap-changer refuses operation.”

a) The procedure of slots filling
As is described above, the slots filling results are shown in Figure 3. The numerical words are extracted by regular expression and they are changed to a standard pattern. That is one drop per minute is translated into one drop every sixty seconds. On the basis of ontology dictionary, all meaningful words are filled in the slots in order.
The 5th International Conference on New Energy and Future Energy System
IOP Conf. Series: Earth and Environmental Science 701 (2021) 012037
doi:10.1088/1755-1315/701/1/012037

Figure 3. An example of slots filling

b) The construction of semantic framework

According to the words’ position information in the initial framework, calculate the distance between each word and fill in the slot according to the construction rules. The match results are shown in Figure 4. Set CS (or A) as the center and compare the distance between each word, thus other slots have the strongest relationship with CS (or A) are determined. For example, while searching for the matched slot of “silica gel”, the value of Dist(“silica gel”, “change color”) is 2 and the value of Dist(“silica gel”, “oil leakage”) is 9. Therefore, the distance between “change color” and “silica gel” is closer and these two terms should match together. The results of semantic framework are shown in Table III.

Figure 4. Semantic framework construction

Table 3. Results of semantic framework

| $C_L$       | Word position | $C_S$      | Word position |
|-------------|---------------|------------|---------------|
| On-load tap-changer | 32            | Radiator   | 13            |
|             |               | Silica gel | 9             |
|             |               | Respirator | 4             |

| $A$        | Word position | $D$        | Word position |
|------------|---------------|------------|---------------|
| Change color | 11            | 60          | 17            |
| Refuse operation | 28            |             |               |
| Oil leakage | 18            |             |               |

c) The combination of words or strings

In Table III, the first and the second line are all same except $C_S$, therefore, these two lines can be combined and the term “respirator silica gel” can be stored in the ontology dictionary. This way to combine the words and strings can not only improve the accuracy of structured data but also make the
dictionary more complete. In practical application, all terms in ontology dictionary are represented by custom code. The final combination results are shown in Table IV.

| $C_L$        | $C_S$               | $A$       | $D$ |
|--------------|---------------------|-----------|-----|
| respirator silica gel | change color       |           |     |
| radiator     | oil leakage         |           | 60  |
| on-load tap-changer | refuses operation  |           |     |

4.2. The statistical analysis of defect results based on text mining

Based on the steps above, 3102 pieces of defect texts have finished the text mining by programming. And in order to verify the accuracy of the mining results, the results are checked manually and the accuracy is about 77%. The reason of inaccurate is that some defect texts are long and redundancy. What’s more, half of the texts are crawled from web and this is another reason that infects the quality of texts. Since the reliability of power equipments’ is high, the number of defect texts still need accumulate. And with lager amount of defect texts, the accuracy of mining the defect texts will become higher. Now the statistics results are shown below according to the mining results of 3102 pieces of defect records.

1) The defect types and quantities of on-load tap-changer in 5 years period are counted, and the results are shown in Figure 5.

![Figure 5. Defect types and quantities of on-load tap-changer](image)

Where, the abscissa represents nine types’ defects of on-load tap-changer. Number 1 represents DC resistance, number 2 represents insulation property, number 3 represents temperature, number 4 represents oil properties, number 5 represents short circuit resistance, number 6 represents aging, number 7 represents mechanical properties, number 8 represents signal communication, and number 9 represents appearance. The vertical ordinate represents the numbers of defects. For example, the defects caused by oil properties have 219 pieces of records.

2) The quantities of each component’s defects in 5 years period are counted, and the results are shown in Figure 6.
Where, component 1 represents the protecting system, component 2 represents the on-load tap-changer, component 3 represents the backstage, component 4 represents the cooler, component 5 represents the transformer body, component 6 represents the winding and component 7 represents the pipe. And the defect types are the same as Figure 5. And as is shown in Figure 6, the oil properties defect of transformer body has the most frequent occurrence.

4.3. The comparison between conventional method and semantic framework
Take the IEC 61970 framework model as the conventional method to make a comparison. And the text samples are 3102 pieces of defect texts in 4.2. After finishing the text mining, the comparison results are shown in Table 5.

| Items                                      | conventional method | semantic framework |
|--------------------------------------------|---------------------|--------------------|
| Accuracy                                   | 71%                 | 77%                |
| Total duration                             | 122 minutes         | 103 minutes        |
| Classification automatically               | No                  | Yes                |
| ontology dictionary                        | No                  | Yes                |
| Percentage of extracting digital information | 70%                 | 100%               |

5. Conclusion
A text mining model based on the modified semantic framework is established in this paper, this provide a new path to realize the information mining for power grid enterprise. And the advantages of this method are as follow, 1) This model is established based on the modified semantic framework and it can process texts in different languages, which is a great improving since the original framework can
only analyse texts in Chinese; 2) Establish the ontology dictionary for the semantic framework is necessary and with a gradually improving dictionary, it will lead to greater efficiency of defect text mining. Since the defect information can be extracted accurately, this method can be applied in defects’ classification and statistics. Moreover, the statistic results can provide basis of establishing the reliability model of the power equipment.

Since the defect text researching in this paper are unstructured data, if the defect information mining from defect texts can be combined with other structured data, such as the equipment condition monitoring data, it can make great contribution to the healthy state assessment, which will become the next main research focus.

References
[1] Yu X 2009 Statics and analysis on reliability of HVDC power transmission systems of state grid corporation of China Power System Technology 33 1-7
[2] Su B L, Yao J G and Du H W 2006 Reliability statistics based on the OPEN-3000 distribute manage system Power System Technology 30 561-6
[3] Rudin C, Waltz D and Anderson R N 2012 Optimal scheduling of composite energy storage for large-scale photovoltaic power generation system based on carbon trading Pattern analysis and machine intelligence 34 328-45
[4] Qiu J, Wang H F and Ying G L 2016 Text mining technique and application of lifecycle condition assessment for circuit breaker Automation of Electric Power Systems 40 107-12
[5] Wang H F, Zhao W F and Du Z D 2015 Economic life prediction of power transformers based on the lifetime data Power System Technology 39 810-6
[6] Deng Y H, Zhang B M and Tong Q F 2001 An approach to automatically generate sequences for 500kV Shuanglong substation Power System Technology 25 30-4
[7] Zhou M, Lin J H and Yang G Z 2004 New-type intelligent dispatching operation order system Automation of Electric Power Systems 28 71-4
[8] Jing C, Chen L S and Wang H F 2017 Semantic framework-based defect text mining technique and application in power grid Power System Technology 41 637-43
[9] Hao X Y, Liu W and Li R 2007 Description systems of the Chinese frame net database and software tools Journal of Chinese Information Processing 21 96-100
[10] Xie J, Shi D Y and Duan X Z 2008 IEC 61970 semantic information model based on ontology technology Power System Technology 32 88-92