Knowledge Extraction Method for Power Grid Fault Text Based on Ontology

Bo Peng¹, Ji Lai¹, Yanru Hao²*, Mengke Yuwen² and Ding Xiao²

¹Information & Telecommunication Company, State Grid Jibei Electric Power Co., Ltd., Beijing 100053, China
²School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China

*Corresponding author email: haoyanru@bupt.edu.cn

Abstract. The power industry usually records equipment failures, defects and other information in the form of text, which contains lots of regular patterns. Knowledge extraction in fault text is of great significance to improve efficiency and reduce the labor cost in the power industry. However, the research for knowledge extraction of text information in this field is rare, it is even more difficult to use machine learning algorithms to mine the deep patterns. To solve this problem, a method of knowledge extraction is proposed in this field. We use power equipment fault texts and relevant guidance as raw materials. Firstly, the knowledge base of this field is designed and constructed based on the ontology concepts, including ontology concept base, description base and regular expression base. Then, the knowledge extraction algorithm is designed according to the knowledge base. After that we conduct the knowledge merge operation to make the extraction results more accurate. Experiments on the real fault texts shows the feasibility and the high accuracy of our method when compared with artificial extraction.

Keywords: Knowledge Extraction; Ontology; Knowledge Base; Fault Text; Transformer.

1. Introduction
In the daily operation and management of equipment in China's power industry, information such as faults, defects and maintenance are often recorded in the form of text. These large amounts of text are accumulated over time and often contain some regular patterns about equipment failures and defects. In the past power grid text analysis, fault texts have strong professionalism and different styles by different writers, so they can often only be performed manually. This not only requires considerable labor costs, but also difficult to mine for some potential regular patterns because of subjective and empirical differences. Therefore, it is necessary to study the mining algorithm for power equipment fault texts. And accurate knowledge extraction of fault information from unstructured raw text data is an important prerequisite for the research of mining algorithms.

Knowledge extraction is to construct knowledge from structured and unstructured raw data. The extraction result is a machine-readable and understandable form, and often has a role in advancing knowledge reasoning. It is not only a reuse of the existing original data, but also a kind of patterned knowledge generation based on the original data [1]. The University of Dortmund in Germany used unsupervised algorithms to extract rules from factual data, and then converted them into PROLOG rules [2]. The Italian SINTESI first performed a full-text analysis, extracting diagnostic knowledge, and integrating a semantic-driven approach based on a general syntactic analysis module [3]. There are many
similar projects, such as the Sabatier project of Paul University in France [4], the SNOWY project of Central University of Florida [5], the KXDC project of Xerox Corporation [6], and so on. It can be seen that knowledge extraction has become a research direction, and the existing main methods all require a large number of detailed regular patterns based on text grammars [7]. Although this can get more accurate extraction results, it will lead to high labor costs and poor algorithm adaptability [8]. Therefore, it is very meaningful to research a set of knowledge extraction methods suitable for power equipment fault texts. In recent years, the research of knowledge graph has gradually become a new hot spot, and knowledge graph is actually an extension of ontology [9]. Both of them emphasize entities and attributes, and ontology more emphasizes the relationship between entities [10-11], which consist many the industry application scenarios that contain professional concepts. So, the idea of knowledge extraction based on ontology structure gradually becomes mainstream [12-13]. However, the research and application of this technology in the field of electric power are still scarce. At present, there is research on reliability of Zhejiang University [14], but the object is the fault texts expressed clearly. For faults with large language style differences and longer length text extraction, it has certain limitations. For fault text, this paper designs an ontology-based power equipment knowledge representation method, and based on the fault case text and equipment condition evaluation and maintenance guidelines as materials, a knowledge base with equipment as the ontology is constructed. In addition, this paper proposes an information extraction method based on the knowledge base, combining the structural information and content information of the knowledge base, which can simultaneously extract the device state quantity and its corresponding description. In the experimental part of this article, the text of the transformer is used as experimental data, and the results of manual extraction are used as a control, which proves the feasibility and high accuracy of the method.

2. Construction of Ontology-based Equipment Fault Knowledge Base

2.1. Ontology Concept and Application
Ontology was originally a concept in the field of philosophy, which was the essence of studying the existence of objective things. Then, the concept of ontology was introduced in the field of artificial intelligence to express the relationship between concepts and concepts in a specific domain or a general domain. In general, the main elements contained in the ontology are concepts, concepts and their synonyms, the relationships between concepts (upper and lower relations, whole and partial relations), attribute relationship of concepts, value range of attributes, etc. [15]. Currently, the concept of ontology has been widely used in many fields, but the application in the field of power is slowly developing, for example, reliability analysis, power grid knowledge retrieval, etc. [16-17] But there are few studies on equipment failure. In this article, we introduce the concept of ontology, which is used for knowledge extraction of fault text.

2.2. Ontology Representation of Power Equipment Information
This article takes the professional concepts in the field of power grids as the design basis of the ontology, and adds consideration to its practical use in engineering applications, describing the concept definition, types, and relationships between the ontology. The relationship between the ontology includes the subordinate relationship, the whole and the part, the synonym and antonym, the attribute relationship of the concept, and so on.
In this paper, we first need to identify the knowledge objects to be extracted from the power grid failure text, and then design the ontology for these knowledge objects [18]. We analyzed more than 600 texts and found that the texts content of power equipment faults contains text descriptions, numbers, symbols, formulas, etc. Since the content is written by different people, the content’s style is not uniform. Combining the actual text data and relevant professional guidelines, we have sorted out the targets that need to be extracted for power equipment fault text as shown in Table 1 below. Obviously, the focus of knowledge extraction is also the difficult part of the component information, which involves the specific structural information of the device. This is also part of the very different description style in the fault text.
Table 1. Targets of knowledge extraction on power equipment fault texts

| Power equipment fault text | Targets                                                                 |
|----------------------------|-------------------------------------------------------------------------|
| Weather condition          | Precipitation, Humidity, Pressure, thunder, Visibility, Wind direction, Temperature. |
| Equipment information      | Manufacture date, Manufacturer, Operation time, Equipment id, Equipment name, Voltage level. |
| Personnel information      | Affiliated company, Name.                                               |
| Oil-height detector        | Description, Parts name.                                               |
| Fault information          | Time, Fault part, Type, Fault classification basis, Fault parts.         |

For the part of equipment, this article proposes to use each type of equipment as the ontology, combining industry guidelines and related professional knowledge to build the ontology. For example, if a transformer is regarded as an ontology, and its equipment information and several components included in it are subordinate concepts of the transformer ontology. The instance below include dry-type transformers, oil-immersed transformers, and fluoride transformers and its components include iron core, winding, insulation, radiator, bushing, thermometer, etc. These components belong to any instance and also belong to the transformer body[18]. We regard a transformer as a ontology, and its instances include dry-type transformers, oil-immersed transformers, and fluoride transformers. The Oil-immersed transformer body, as shown in Table 2 below (because there are many components in transformer ontology, here is mainly shown the oil-immersed transformer):

Table 2. Ontology components’ conceptions of transformer

| Oil-immersed transformer’s components | Components                                               |
|--------------------------------------|----------------------------------------------------------|
| Body                                 | Iron core, Winding, Lead, Insulation, Conductive joint.   |
| Tank                                 | Throttle gate valve, Respirator, Oil conservator, Body.   |
| Cooling device                       | Oil pump, Fan, Heat sink.                                |
| Protective device                    | Thermometer, Current Transformer, Gas relay, Explosion-proof valve, Pressure relief valve, Oil flow relay. |
| Outlet device                        | Casing pipe.                                             |

It can be seen from Table 2, the higher the concept degree, the more sub-concepts it contains. In theory, the conceptual objects for description are the component of the body, that is, the most specific unambiguous parts. But in actual text, it may also involve describing high-level concepts. Therefore, for the construction of the device ontology concept base in accordance with the hierarchy, and the more detailed the more it can reduce the information loss caused by the incomplete matching. At the same time, because different writers may use different aliases for the concepts in it, for example, the main transformer “zhubianyqi” may be called “zhubian”, and the oil filter “youlvji” may be called “ivyouzhuangzhi”, “jingyouqi”, “jingyouxiang” etc. So, in order to be able to better suit the actual situation, we also need to build a subsidiary attribute for each concept, that is, its synonyms.

2.3. Design of Ontology-based Power Equipment Knowledge Base

With the knowledge of the transformer structure, it is possible to complete the extraction of component keywords for knowledge extraction, but it is not yet possible to extract the content of component descriptions. Due to the strong professionalism of power grid failure texts, the description terms of components are standardized and unified. Therefore, it is feasible to match descriptions by organizing and summarizing the description knowledge of components. We combined more than 600 fault texts and related guidelines to complete the description of the nature of the transformer components. The partial diagram is shown in Table 3 below (due to space limitations, here are mainly shown some components of oil-immersed transformers):
Table 3. Property description base of transformer

| Fuel tank ontology components | Property description                                      |
|------------------------------|----------------------------------------------------------|
| Radiator                     | Filthy, Rust.                                            |
| Lead clip                    | Damage, Loose.                                           |
| Conductive joint             | Heat, Loose.                                             |
| On-load tap-changer          | Counter failure, Refuse to move, Slip, Air switch does not close, Internal oil leakage, Unusual sound inside. |
| Casing pipe                  | Filthy, Damage, Heat, Oil seepage, Surface anomaly, Discharge, Crack. |

It can be seen from Table 3 that the organized description can comprehensively cover the nature of the corresponding components. However, in the actual text, there are many descriptions of the degree, such as "severe pollution", "high oil level". In order to be able to better extract various information in the text, the composition of the description base needs to be divided into two parts: property description and degree description. Property are descriptions of conceptual states such as "oil spills" and "dirty"; degree describe the degree of states such as "seriously", "slight", and "over". Such an organization method can adapt to a wider description form through permutation and combination of two types of descriptions in the subsequent extraction. At the same time, it can achieve a more accurate level of the state of the extraction.

2.4. Design Advantages of Knowledge Base

Through the construction of the component structure relationship and the completion of the description base, the fault knowledge base for the equipment is completed. Fault knowledge representation based on ontology not only lays a good foundation for subsequent knowledge extraction, but also facilitates the maintenance and modification of the knowledge base. For example, the addition, deletion, and modification of component descriptions only need to perform corresponding operations on the corresponding entries in the description base of the component. That can greatly reduce errors caused by operational errors, and the modification operations of components is the same.

3. Structured Knowledge Extraction of Text Information

3.1. Overview of the Extraction Process

The goal of knowledge extraction for power equipment fault text is as shown in Figure 1. The extracted description can be divided into numeric symbol types, text type and mixed type. For example, the description of the device model is a numeric symbol type, the description of the heat sink is a text type such as "rust" "seriously contaminated", the description of the total hydrocarbon is mixed type consisting of two forms such as "210.7ppm" and "substantially rising". Because different extraction methods are used for different types, in the construction of the ontology knowledge base in section 1.3, the description type of the concept needs to be added to the knowledge base as a feature of the concept. In some knowledge-extracting literatures, sentence dependency analysis often performs first, and obtain the parts of speech of different words. Then parts of speech are used as labels to correspond to different types of semantic slots. The steps of knowledge extraction designed in this paper are mainly divided into two steps, as shown in Fig 1 below. For each text sentence, first extract the concepts contained in the device's ontology base, and then for each concept, determine which of the three types it belongs to. If it is a number symbol type or mixed type, combined with a regular expression base for extraction; if it is textual and mixed, extract it in combination with the description base.
3.2. Merge Results

After the process of the previous steps, fault information of the form <concept, description> can be extracted from the text. But there will be a lot of redundant information in this, because when extracting the ontology concept, only all the concepts are directly extracted. Without considering the relationship between concepts, some subordinate concepts can actually be combined and reduced to the lowest level concepts. For example, a fault sentence is "It is found that the 220kV # 2 main transformer on-load tap-changer counter is damaged, the oil tank is leaking seriously, and the oil level gauge is damaged."

It’s results of the extraction process after 2.1 are shown in Table 4:

| Concept              | Properties         | Degree   |
|----------------------|--------------------|----------|
| voltage              | 220kV              |          |
| on-load tap-changer  | damage             |          |
| counter              | damage             |          |
| tank                 | oil leakage        | seriously|
| oil-height detector  | damage             |          |

The text here is actually expected to express <on-load tap-changer counter, damage> instead of <on-load tap-changer, damage> and <counter, damage> two pieces of information that are redundant or even make misunderstanding. Therefore, the step of knowledge merging should be performed after extraction. The main steps of knowledge merger are shown in Fig 2 below.
Two identical extraction results: concepts, description:

The parts of “concept” are exactly the same

The parts of “concept” have the relationship of subordination

Determine whether the merger conditions are met

Merge

\[
\text{strcat} (\text{Parent concept, Child concept}), \text{description} \leq (\text{<Parent concept, description >,}
\text{<Child concept, description >})
\]

**Figure 2.** Process of knowledge combination

It is necessary to add this step of knowledge merger after knowledge extraction. After the combination of the two operations, not only complete the extraction from the content in the extraction step, but also consider of the equipment structure in the step of knowledge integration. Such considerations are rarely reflected in previous methods of knowledge extraction.

4. **Experimental Process and Experimental Analysis**

In order to prove the validity of this method, we take the relevant standard guidelines for transformers and 627 transformer fault texts as experimental data. We first constructed the overall knowledge base of the transformer equipment through summary guidelines and 274 texts. The composition structure is shown in Table 5 below.

**Table 5.** Property description base of transformer

| Knowledge base          | Contents                                                                 |
|------------------------|--------------------------------------------------------------------------|
| Regular expression base| Description                  | regular expression |
|                        | Short-circuit current          | \{[0-9]+\}.\{[0-9]+\}\{[0-9]+\}+Set\{A,mA,kA\} |
|                        | ...                           | ... |
| Synonym table          | Iron core                   | Iron heart |
|                        | Oil pillow                  | Oil conservator |
|                        | ...                           | ... |
| Description base       | Property description         | Level |
|                        | Heat sink                   | Serious |
|                        | Rust, filthy,……            | ... |
|                        | Oil pillow                  | Slight |
|                        | Oil leak, rust, High oil level, low oil level, … | ... |
As mentioned in Sections 1 and 2, it includes the concept of ontology, that is, the relationship network between various concepts, the synonyms of each concept, and the description category corresponding to each concept. The description base, which is a collection of literal and mixed concept descriptions. There are two types of descriptions: nature description and degree description. Regular expression base is the collection of numeric symbolic concept descriptions; whose content is the regular expression corresponding to each concept.

For the evaluation of the extraction results, the artificially extracted <concept, description> results were used as a control group for the program extraction results. Because there may be multiple descriptions of a concept in the text, and the extraction result may not completely hit all the descriptions, there will be a lot of errors if directly comparing the results of <concept, corresponding description set> for each concept. Therefore, the results are judged in terms of concepts and descriptions. Suppose that the concept of artificially extracted knowledge is \( T_{N_c} \), and the concept is defined by the program as \( P_{N_c} \). If the two overlap, the correctly recognized concept is \( T_{P_c} \). Similarly, describe the corresponding manual extraction is \( T_{N_d} \), the program is regarded as \( P_{N_d} \), and the program correctly is regarded as \( T_{P_d} \). The three evaluation indicators precision, recall, and \( F_1 \), are defined as follows (where \( i \) is the same as \( c \) or \( d \)):

\[
\text{precision}_i = \frac{T_{P_i}}{P_{N_i}} \times 100\%
\]

\[
\text{recall}_i = \frac{T_{P_i}}{T_{N_i}} \times 100\%
\]

\[
F_1 = \frac{2 \times \text{precision}_i \times \text{recall}_i}{\text{precision}_i + \text{recall}_i}
\]

We performed extraction experiments on the remaining 353 texts, and the results are shown in Table 6:

**Table 6. Results of knowledge extraction experiment**

| Concept | PN before integration | PN after integration | TN     | TP     | Accuracy Rate | Recall Rate | F1 |
|---------|-----------------------|----------------------|--------|--------|---------------|-------------|----|
| Concept | 5304                  | 4256                 | 4613   | 3876   | 91.1%         | 84.0%       | 87.4% |
| Description | 6377                 | 5327                 | 6775   | 3998   | 75.1%         | 59.0%       | 77.1% |

It can be seen from Table 6 that the operation of knowledge merge is very necessary, which can reduce redundant concepts and descriptions by about 24% and 19%, respectively. So, the accuracy rate will be improved and the extraction results will be more accurate. For the concept extraction, because the concept of power equipment is relatively clear, almost all the programs can extract hits, and the accuracy rate can reach 91%. A sentence may contain multiple concepts in the text description, but for some of them, there is no corresponding description of the concept, and its emergence may only be used to explain the location of other concepts. But according to the extraction algorithm, the program will also extract it, such as the concept of "bushing" in "there is a discharge signal at the main transformer body below phase C of the 110kV high-voltage bushing of the No. 1 main transformer" is to explain the specific fault location of the transformer body. However, it will be extracted as a concept and the corresponding description cannot be found, which is one of the main reasons to results in recall rate lower than accuracy rate. However, in general, the concept extraction has achieved relatively satisfactory results. As for the description extraction, the result is slightly worse than the concept extraction. It is mainly because the description in the actual text is more flexible and diverse. It can be improved by further complementing the description base. Since the description base is constructed using only 43% of the text data, but the results have already reached 66% of the measure of \( F_1 \), so further completion through more fault texts, there is still space for improvement.
5. Conclusion
This paper mainly researches and implements the knowledge extraction method for power equipment fault texts, and proposes a set of ontology-based domain knowledge base construction ideas, including equipment ontology concept base, description base and regular base for knowledge extraction. Combine the preliminary extraction results through the equipment ontology concept base to improve the accuracy of the results. Summarizing the results of this article, the main points are as follows:
1) Design and build the power equipment ontology concept base, concept corresponding description base, and digital symbol type description regular base. The design of the knowledge base is continuously refined by classification, making the granularity small. It can be arranged by matching the combination method adapts to a wide range of expressions, so as to perform accurate and complete extraction;
2) Design the algorithm of knowledge extraction and knowledge merge based on knowledge base;
3) Through the experiments, the feasibility and high accuracy of the extraction process have been proved;
4) Through the analysis of experimental results, the improvement direction of the process is proposed. In the future, we decide to improve the information in the table through design rules.

Acknowledgements
This paper is funded by the State Grid Jibei Power Co. Ltd. Science and Technology Project “Research on Data Fusion and Visualization Key Technologies Based on Full-service Unified Data Center” (52018E18006N).

References
[1] Jedrzej Potoniec, Dawid Wiśniewski, Agnieszka Ławrynowicz, C. Maria Keet. Dataset of ontology competency questions to SPARQL-OWL queries translations[J]. Data in Brief,2020,29.
[2] Ultsch A. Knowledge Extraction from Self-Organizing Neural Networks[M]// Information and Classification. Springer Berlin Heidelberg, 1993:301-306.
[3] Alani H, Kim S, Millard D E, et al. Automatic Ontology-based Knowledge Extraction and Tailored Biography Generation from the Web[J] // IEEE Intelligent Systems, 2002, 18(1):14-21.
[4] H. Alani,,S. Kim,,D. Millard et al. Automatic Ontology based Knowledge Extraction from Web Documents[J]. IEEE Intelligent Systems. 2003,18(1):14-21
[5] Ciravegna F, Campia P, Colognese A. Knowledge extraction from texts by SINTESI[C]// Conference on Computational Linguistics. Association for Computational Linguistics, 1992:1244–1248.
[6] Sbattella L, Tedesco R. Knowledge Extraction from Natural Language Processing[M]// Methodologies and Technologies for Networked Enterprises. 2012:193-219.
[7] Hamed Karimi,Ali Kamandi. A learning-based ontology alignment approach using inductive logic programming[J]. Expert Systems With Applications,2019,125.
[8] ZHANG Zhixiong, WU Zhenxin, Liu Jianhua, et al. Analysis of State-of-the-Art Knowledge Extraction Technologies[J]. New Technology of Library and Information Service, 2008,24(8):2-11
[9] M. Gayathri,R. Jagadeesh Kannan. Ontology based Concept Extraction and Classification of Ayurvedic Documents[J]. Procedia Computer Science,2020,172.
[10] HU Fanghui. Chinese knowledge graph construction method based on multiple data sources [D]. East China University of Science and Technology, 2015
[11] LIU Qiao, LI Yang, DUAN Hong, et al. Knowledge graph construction techniques[J]. Journal of Computer Research and Development, 2016,53(3):582-600
[12] CHEN Jing. Research of Ontology-Based Information Extraction[D]. Soochow University, 2007
[13] WANG Xiangqian, ZHANG Baorong, LI Zonghui, Overview of Ontology Research[J]. Journal of Intelligence, 2016,35(6):163-170
[14] Qiu Jian. Research on Chinese Textual Data Mining Techniques and Reliability Applications in Power Systems[D]. Zhejiang University, 2016
[15] DU Xiaoyong, LI Man, WANG Shan. A survey on ontology learning research[J]. Journal of software, 2006, 17(9): 1837-1847

[16] YANG Yuehua, DU Junping, Ping Yuan. Ontology-Based Intelligent Information Retrieval System[J]. Journal of Software, 2015, 26(7): 1675-1687

[17] LI Yang. The knowledge retrieval system of smart grid based on ontology[D]. Beijing Institute of Technology, 2011

[18] XIE Jun, SHI Dongyuan, DUAN Xianzhong. IEC 61970 semantic information model based on ontology technology[J]. Power System Technology, 2008, 32(1): 88-92.

[19] Gaha M, Zinflou A, Langheit C, et al. An Ontology-Based Reasoning Approach for Electric Power Utilities[C]// International Conference on Web Reasoning and Rule Systems. 2013: 95-108.