ENMFM Algorithm Applicable for the Multi-Source Heterogeneous Data Source Model of Power Grid Regulatory Cloud Platform

Lan Haibo¹, Xiao Linpeng²,³ *, Mu Yongzheng¹, Cao Liangjing¹, Liu Huiyong²,³, Lou Tianyue²,³

¹State Grid Jibei Electric Power Company Limited, Beijing 100053, China
²Nari Group Corporation/State Grid Electric Power Research Institute
³BeiJing Kedong Electric Power Control System Co., Ltd, Beijing, 100192, China
*524869920@qq.com

Abstract—With the rapid development of the intelligent grid technology and the arrival of the informationalized big data era, the application of computing method to power grid specialty will be a key step for the value of power grid data. Grid data record is not standard, the confusion caused by a device with multiple recording, professional equipment for power grid name recognition and use of data and algorithm based on identification service work will gradually replace artificial recognition, this paper presents a ENMTM algorithm with the grid raw data, which can identify accurately the same equipment with different names. The equipment name record is an important data of power grid business scenario, however, the same device is easy to describe in different situations or staff’s records, which causes data redundancy and equipment confusion caused by error identification of equipment name. The name recognition is to segment the device name string into a structured element unit which can be calculated by the computer, so as to realize the identification of multiple words with one meaning and devices with different names. In this paper, we study the recognition algorithm of recognition, the establishment of professional word library, segmentation and calculation of similarity recognition. The results show that the ENMFM result error rate is only 3.13%, on the basis of segmentation results and the confusion matrix of device name similarity can be seen that the positive samples has good prediction. In addition, in the experiment, the Top N items of higher similarity were given, which provided the staff with professional judgment to ensure the rigor of work.

The calculation method to solve the name recognition of power equipment can greatly reduce the human work of identifying equipment, improve the work efficiency of power professional business scenarios, and lay a solid foundation for the unified data environment of intelligent grid.

1. INTRODUCTION

As the construction of electric power grid develops, new generation of regulatory system integrates the emerging Internet technologies step by step and constantly undertakes infrastructural and architectural upgrade towards the direction of “cloud calculation, big data, the Internet of things and mobile Internet”. Among them, the subject of building a regulatory cloud platform may have a style of itself. Taking on
the significant mission of integral system construction under new era and new concept, it combines big data, micro-service and other cutting-edge technologies, as well as multiple classic regulatory management systems in use to form the public cloud platform and private cloud platform of electric power grid dispatching itself.

However, the naming criterion of electric power equipment is absent from unified and standardized management, so the multiple classic systems in use possess their respective equipment naming standard. For instance, the D5000 system uses five-section naming style; the OMS system uses optional three-section naming style; and the PMS system integrates the characteristics of D5000 and OMS. Both above naming styles appeared in PMS system variably. What’s worse, some autonomous dispatching management systems use and even maintain a collection of model account information in accordance with their own business demands. In addition, the cloud platform itself also possesses a set of naming conventions that has big difference from above systems. These issues lead to a great amount of conflicting cases when the cloud platform combines with the equipment account model that was involved in above systems. Currently the only solution is business staff revises the information or compile generic SQL manually. Thereupon it brings about extremely low efficiency in matching and revising the account model of multi-source heterogeneous system equipment as well as reducing the compatibility and expansibility of application. Accordingly, this paper indicates that, it will be a high priority and major difficulties in promoting the construction of power dispatching cloud platform along with the intellectualization of electric power dispatching by resolving the semantic information of equipment names, identifying and matching the basically same equipment account that possesses various names in accordance with its characteristic attributes.

2. OVERVIEW OF RESEARCH STATUS

Verified by the references, as of now researches on similarity matching of equipment names have been launched within both power grid industry and related industries. Mostly these researches were developed in view of the editing distance differences by text comparison. They used the LD similarity algorithm and jarccard similarity algorithm to match and configure the similarity degree of database formulation column. To the comparative table of item output that exceeds the threshold value, it can identify the mismatching of common naming. However, as to the algorithm that merely compares the editing distances, in that the equipment names were handled as two given texts, the matching work was dependent on the segmentation sequence of character string. Therefore, such algorithm is incapable of recognizing the non-standard equipment name or the names that missed record. Furthermore, it is unable to provide favorable solution [2] for professional word library, simple Chinese or English naming in professional power grid business scenario.

3. CHARACTERISTIC EXPRESSION AND NORMALIZATION OF ELECTRICAL EQUIPMENT NAMES

3.1 Characteristic Expression of Electrical Equipment Names

Nowadays, it is imperative to put forward unified and standardized management on the naming criterion of power grid equipment. When the system application is dispatched, one or more sets of account information are frequently possessed and maintained by various applications. The same equipment has multiple expressions in different business systems, specifically to the same section, every staff has divergent expressions.

Typically, the electrical equipment name is composed of continual characters and numbers. Even if the delimiter is not counted in, the sequence of string with characters and numbers varies from one another. In the process of segmentation, it may occur to various interpretations; and after segmentation, the character string may contain complicated cases of sequencing. In comparison with dealing with pure Chinese or English names, it is more complex to identify this kind of equipment name. Additionally, within the system the geographic name and address name are absent from standardized management for a long time. The abbreviation records and else multiple expressions that were taken with local and
personal habits also aggravated the difficulty in equipment name recognition. For instance, “Mazhuang 220kV #4 Female cable” was recorded as “Mazhuang4F”.

3.2 Demand on Normalized Expressions of Electrical Equipment
Due to above mentioned variability and irregularity in manual naming, it brought about the problem that the same equipment possesses different expressions. Accordingly, this interferes with real-time processing of warming signal and other abnormal information. Further in terms of industrial efficiency, in China both residential and industrial power supply live with power grid. With an increasing impact on the dispatching system, people call for this kind of normalization on equipment naming. That is to say, if the semantic information in equipment names can be resolved accurately, it can recognize different naming which is supposed to be the same equipment as well as fulfill the unified and standardized management. It is of great significance to improve the operational efficiency of the interactive industry between power and Internet.

3.3 Difficulties and Key Technologies of Electrical Equipment Name Recognition
The difficulties in equipment name recognition consist of following aspects: the first issue is establishment and storage of professional word library. As a professional field that involves physics, communication and more other subjects, in electric power industry, the names of power grid definitely consist of professional industry vocabulary. Many of them are combination of geographic names and else regular pre-corpuses, thus forming synthetic and complex names; secondly, in contrast with handling pure Chinese or English names, it is more intricate to recognize the equipment name that was made up of continual characters and numbers. Without delimiter, it has various sort orders and interpretations in the midst of segmentation. In addition, the early historical data in system were recorded a long time ago, so the geographic name and address name were not kept as per standard. Otherwise, the staff recorded the abbreviations and else multiple expressions in accordance with their personal habits. This also aggravated the difficulty of equipment name recognition.

To solve above problems, this paper probed into the data value of electric power. Moreover, it used the ENMFM data structure to deposit the vocabulary of electric power, and built the lightweight word library that accorded with the business scenario. The word was segmented accurately in the light of professional word library; through acquaintance level recognition algorithm, it resolved the semantics of equipment descriptions so as to form structured object that could be processed by computer. Consequently, it accurately captured the real intention behind the descriptions of system equipment, thus realizing the equipment matching of acquaintance.

4. ENMFM ALGORITHM

4.1 The Conformation of Electric Power Term Dictionary
Professional power equipment names are technical terms which are constituted as per certain rules. In the meanwhile, these names possess few common prefixes. To solve the problem that arises from these features, the model is designed to deposit the equipment names in a special tree-structure, i.e. Ac automation, thus setting up the specialized electric power vocabulary.

Ac automation is designed as a tire tree that is equipped with a fail needle. In case the matching of current point fails, the needle will be transferred to the location where fail needle points at. In this way, backtracking will be saved to streamline the matching smoothly. The suffix of current pattern string is same as part of the pattern string prefix that the fail needle points at. For “220kV Changzhi Main Transformer” and “Changzhi Transformer” for example, when the word “Zhi” is oriented and its next word is not the word “Main” we are looking for, it will orient to the word “Transformer” in “Changzhi Transformer” string. Then it will check if the next character “Transformer” is the target. This method integrates the efficiency advantage of digital search tree, and the special advantage of binary search tree. In summary, it is appropriate to resolve the professional vocabulary issues [3].
4.2 Segmentation of Chinese Words

Regularization of Text

To date, the habits that record the equipment names vary individually in the electric power system throughout the country. Its present situation is very serious for different expressions on the same equipment. Thereupon, before the original names of equipment are standardized, in addition to the code processing, the non-Chinese character filtering and else pretreatment work, a top priority is unifying the writing of numbers. For instance, to “Mazhuang 220kV #4 female cable”, it has a couple of common expressions: “Mazhuang 220kV No.4 female cable”, “Mazhuang 220kV #4 female cable” and “Mazhuang 220kV No. 4 female cable”. In this paper, it used regular method to normalize all the numbers in the equipment names, so as to convert them into unified Arabic numerals.

Word Segmentation

The equipment expression describes the implementation of its definition. It is a given label for specific equipment. The real intention is distinguishing this device from other devices. The working staff can understand the meaning of equipment expression and abstract the latent information because they have knowledge of the technical information of this device and its conventional description. To comprehend the semantic information of the equipment expression, the computer also needs to learn the relevant knowledge of name resolution. It is unable to have integral knowledge and comprehension of the whole input character string. Therefore, it is necessary to segment the Chinese words. The character string of equipment name is converted to segmented arrays that can be recognized successively by the computer.

A major type of Chinese word segmentation algorithm is developed by character string matching. That is scanning the character string. Once the substring in the character string is identified to be the same as the word, the system will consider it as matching [4]. Normally some heuristic rules will be added in, such as “the maximum matching in forward direction/reverse direction”, “Long word first” and else strategies. The matching goes fast and it is all about O(n) time complexities [5].

In case the ambiguity and unknown words are not handled properly in electric power field. For example, the word “photovoltaic power station” in the expression “Tengyuan Huichuan photovoltaic power station 220kV #5 female wire” may never appear, so the granularity of word segmentation is also an index to be considered. This paper fulfilled the maximum matching results of word segmentation in accordance with the optimized dictionary. Therefore, it could avoid such problems. The algorithm was described in Algorithm 2.
4.3 Matching of Semantic Similarity

Nowadays the habits that record the equipment names vary individually in the electric power system throughout the country. Its present situation is very serious for different expressions on the same equipment. For instance, to “Mazhuang 220kV #4 female cable”, it has several common expressions as follows: “Mazhuang 220kV No.4 female cable”, “Mazhuang 220kV #4 female cable” and “Mazhuang 220kV No. 4 female cable”. Therefore, after the segmentation of original equipment expressions, it calculates the similarity in view of the effect of word segmentation. Moreover, before taking into account of the similarity among the standardized words, it is supposed to consider the word order as well as the similarity between the Arabic numerals and special characters.

This paper used the similarity algorithm in calculation on the basis of semantics and word order. The regular method was used to standardize the numerical matching in equipment names. Then these numbers were converted into unified Arabic numerals; In the light of given equipment expression, after it was segmented through the ternary search dictionary tree, the equipment vectors including all words were obtained. Next, the redundant strings were removed before merging into the general vector which was expressed as T; Then in accordance with the semantic vector and formula, the similarity $S_s$ was calculated. The similarity of word order was obtained by calculating the semantic similarity and the word order similarity between two sentences, and then it used the weighted method to get the similarity of these two sentences. At last, the semantic vector cosine formula was used to get the similarity $S_{s_s}$ of the equipment naming. The algorithm was shown in Algorithm 3.
5. ANALYSIS OF NUMERICAL EXAMPLES

In this paper, the JDK 1.8.0 was selected as the developing platform. The Dameng 7 was selected as the database platform. The software system of electric equipment characteristic matching was developed on the Ningsi 6.8.0 operation system.

5.1 Construction of Data Machine

To verify the validity of the method mentioned in this paper, 3270 pieces of data were selected as the experimental data from the real equipment names of various business sections in electric power system. On this basis, two data sets were constructed respectively for word segmentation test and similarity matching test: (1). Sampling 450 pieces of equipment name testing set at random as the data set of word segmentation experiment; (2). Sampling 100 pieces of equipment names at random as the positive samples. Otherwise, 100 pieces of fake equipment names were prepared manually as the negative samples. In them the correct word segmentation results and the similarity matching results were marked out, thus accurately evaluating the two important intermediate results in the process of electric equipment recognition.

5.2 Evaluation Index

The segmentation of electric equipment names is an essential part in recognizing the same equipment that possesses different names because it directly affects the results of equipment recognition. In this paper, it refers to the common evaluation criteria in Chinese word segmentation, namely it uses the precision, recall rate and error rate to evaluate the results of word segmentation. The formula was shown as follows:

$$\text{Precision} = \frac{TP}{TP + FN}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Error Rate} = \frac{FN}{N}$$

TP represents the number of words which are segmented correctly; FN represents the number of words which are wrongly segmented; N refers to the number of words which are segmented as per golden standard. In them, TP and FN represent the numbers of difference in word segmentation. Accordingly, it sorts out the number of words which are segmented correctly and wrongly respectively.
### 5.3 Segmentation of Chinese Word

This paper drew 3270 pieces of data those were currently used by the power plant. These data were then used to build the data set of word library, and the ternary dictionary search tree. It laid an efficient foundation for the segmentation of equipment names. Again, 450 pieces of equipment name testing sets were drawn at random. In the experiment, for instance, “Langfang north China national grid north Hebei maintenance company” accorded with the golden standard of segmentation. “Langfang north China national grid north Hebei maintenance company” was the result to be evaluated. Each segmented word had a location tag. (1,3), (3,6), (6,8) and (8,12) were golden standards; (1,3), (3,6), (6,8), (8,10) and (10,12) were the results to be evaluated. By comparing the number of variabilities in both tagging results, it obtained the quantity of words those were correctly marked and those were wrongly marked respectively.

The experimental results of equipment name segmentation were shown in Table 1. It showed that, the error rate of word segmentation by using ternary search tree was only 3.13%. It could be seen that, the ternary search tree and the forward longest matching algorithm played an important role in building the professional word library.

#### Table 1 Word Segmentation Results

| Conditions of word segmentation | Precision | Recall | Error Rate |
|--------------------------------|-----------|--------|------------|
| Word library of dictionary tree | 96.82%    | 95.33% | 3.3%       |
| Normal dictionary              | 76.86%    | 90.11% | 27.13%     |

### 5.4 Similarity Matching

3270 pieces of data were used to build the data set of word library. Consequently, similarity calculation was carried on the basis of word segmentation. The data set in use comprised 100 pieces of positive data samples those were abstracted from the 3270 pieces of data at random, as well as 100 pieces of negative samples those possessed unknown power plant equipment. In that the result of equipment name similarity was specific decimal, therefore, this paper set the threshold value as 0.88 to be the false positive standard in accordance with the experimental results. That was, if the similarity level of predictive negative samples was more than 0.88, it matched the prediction. The results of confusion matrix in the equipment name similarity matching test were shown in Table 2. From this table, it can be seen that the positive samples were perfectly predicted, accordingly proving the effect of word library construction and the word segmentation; on the other hand, given the threshold value was set as 0.88, the proportion of negative samples those were predicted as positive was 44%. This was because, in terms of the problems under professional business scenarios of power plant, the experiment required the first N equipment which possessed relatively higher similarity level. The equipment was provided to the working staff for professional manual judgment and thus ensured the accuracy of work. The threshold value in this experiment was designed for this purpose.

#### Table 2 Similarity Matching Results

| Label/predict | Negative | Positive |
|---------------|----------|----------|
| Negative      | 44       | 56       |
| Positive      | 0        | 100      |

### 6. Analysis of Numerical Examples

The same equipment developed various journaling styles in different scenarios or regions. This brought about huge difficulties and thus caused confusion in professional scenarios of equipment recognition. The sub-standard recording of equipment names exists in different business sections and working scenarios of electric power company. It has always been a major issue in the midst of equipment management and data analysis. Manpower can be greatly saved for recognizing the equipment if this
issue is settled down. What’s more, the operational efficiency under professional business scenarios will be promoted, consequently laying a solid foundation for the unified data environment under intelligent power grid.

This paper aimed to set up a professional word library for the professional business scenarios of electric power field. On this basis, it provided an effective and targeted data base for equipment recognition. By integrating the similarity matching algorithm based on semantics and word order, it fulfilled the accurate matching of equipment names. Shown from the experiment, it was a workable and effective method. On the other hand, the limitation of this method in the paper is that, it is subject to the constant growth of present electric power professional vocabulary. The established dictionary is requested to be maintained in accordance with the professional word library. Therefore, in future study, it will focus on how to efficiently maximize the expansion and maintenance of the professional dictionary

REFERENCES
[1] Ma Qiang, Guan Jia, Bai Yingwei, Wang Zhaoxin, Liu Yong. Research on Equipment Association Technology of power grid business system based on big data [J]. Automation and instrumentation, 2018 (04): 10-13 + 17
[2] Wang Fang, Wang Jirong, Yang Xiaodong, Li Jun. improvement of editing distance algorithm based on Chinese text [J]. Journal of Qingdao University (NATURAL SCIENCE EDITION), 2017, 30 (03): 60-63
[3] Xu Yibin, Xu Xuerong. A new ordered data structure: capacity balanced Trident search tree [J]. Computer engineering and applications, 2017, 53 (09): 63-71 + 102.
[4] Tong Xiaoguang, Guo Chao, Kang Hongjing. Optimization of short Chinese word segmentation algorithm [J]. Software guide, 2017, 16 (05): 42-44.
[5] Chen Minghua, Yin Jinghua, Shu Chang, Wang Mingjiang. Implementation of word segmentation system based on forward and reverse maximum matching [J]. Information technology, 2009, 33 (06): 124-127