Ternary Search Trie based Algorithms for Recognizing the Names of Power Devices

Yi Liu¹, Jiahui Xu ²,³, Xiujun Gong ⁴, Jingjie Bai ²,³, Linpeng Xiao ²,³, Hongqin Zhu⁵, Chen Tan ²,³, Tianyue Lou ²,³

¹State Grid Jiangsu Electric Power Co., Ltd., Nanjing 210029, China
²Nari Group Corporation/State Grid Electric Power Research Institute, NanJing 211000, China
³Beijing KeDong Electric Power Control System Co., Ltd., Beijing 100192, China
⁴School of Computer Science and Technology, Tianjin University, Tianjin 300072, China
⁵State Grid Nanjing Power Supply Company, Nanjing 210019, China

Abstract. With the development of the smart grid technology and the advent of the big data, it is highly desired to develop efficient computational methods for mining the values behind of huge volumes of data on the grid. However, non-unified naming of the electronic devices from the de-centralized and heterogeneous power systems makes the procedure a hard problem. In the present work, we proposed a Ternary Search Trie based algorithm for recognizing the devices from the huge volumes of transaction data in the power grid. On the base of well-defined professional vocabulary of electronic devices, the algorithm scans and segments records having names of devices using the Ternary Search Trie method. An integrative scoring schema combing semantic and word order similarity is designed to evaluate the segments against the target device names. Experimental results indicate that the proposed method achieves only 3.13% error in parsing stage, and higher recognition rate. Moreover, the algorithm lists top-n matches for professional judgments to make the work more useful and rigorous.

1. Overview
Stable and reliable power equipment is the basis and guarantee for the safe operation of the power grid [1]. With the development of smart grid technology and the integration of ultra-high voltage power grids, new energy grids and micro-grids, the types and quantities of power equipment are growing rapidly [2]. At the same time, the information platforms at all levels of the power grid, such as monitoring, dispatching and overhaul, have accumulated a large amount of equipment operation and maintenance data, which can provide intelligent decision support for scientific assessment, accurate warning and targeted maintenance of equipment defects and faults [3, 4, 5]. However, due to the heterogeneity of information systems of maintenance and operation at all levels of the power grid and difference in automation degree (some information still needs to be manually filled out), the device names are different in different systems, which seriously hinders the data use efficiency [6]. Therefore, the development of efficient and accurate device name identification method is the basis and premise for the effective use of large data grid.
Power device name recognition is essentially a category of Named Entity Recognition in natural language processing, mainly including grammar-based methods and machine learning-based methods [7]. The former requires a wealth of grammar rules and dictionaries, while the latter requires a large number of corpora for training [8, 9].

State Grid Corporation released the naming specification for general data model of power grid equipment ("DL/T 1171-2012") in 2012 and partially revised it in 2017 ("GB/T 33601-2017"). It gives a complete expression of the full path name of the grid equipment, that is, the name of the power equipment is generally composed of four parts: grid name, voltage level, equipment number and device semantic name to which the equipment belongs. However, the combination diversity of various expressions of the parts makes the traditional named entity recognition method not directly applicable to the identification of power device names.

This paper firstly builds a standardized device name database according to the general data naming specification of grid equipment developed by State Grid Corporation, and establishes a corresponding name dictionary based on tri-fold search tree. Secondly, for the operation and maintenance information records that may include the device name, such as the maintenance ticket, the description of the alarm information, etc., this paper performs word segmentation, constructs corresponding word order and semantic representation vector. Finally, this paper identifies the target device by calculating the semantic and word order similarity of the record with the device name in the standard database.

The methods suitable for Chinese word similarity measurement mainly include the method of editing distance [10] and Jaccard method [11]. The former uses the dynamic programming method to calculate the number of edits between two strings, which is suitable for comparison between pure device names without interference of other text. But for the overhauled text with semantics before and after, the matching accuracy is greatly reduced due to the length of the text and the position of the text to be recognized. The latter counts the number of shared elements between the two strings to be matched and the number of elements, which is suitable for comparison between pure device names without interference of other text. But it is susceptible to semantic interference in the repaired text and weak in recognizing continuous words and digital combinations. So it fails to provide solutions for the professional vocabulary and simple Chinese or English naming faced by the professional grid business scenarios.

Based on the in-depth analysis of the characteristics of power equipment names in the operation and maintenance systems of power grids, this paper establishes a power equipment model with place names, numbers and device semantic names, and builds a device name dictionary library based on the tri-fold dictionary search tree. This structure combines the speed advantage of the digital search tree with the space advantage of the binary search tree, making it easy to expand and maintain. Based on this, an algorithm for device name word segmentation and a similarity calculation method based on semantics and word order are developed.

2. Characteristics of Power Equipment Name Expression
Due to the heterogeneity of the operation and maintenance information systems at all levels of the power grid and difference of automation (some information still needs to be manually filled out), there is still a long time to go to popularize grid equipment naming standardization. On the one hand, the information systems of the power grid need to integrate the original data, but most of these systems lack standardized management of device naming. On the other hand, some forms of the scheduling system still need to be manually filled, but personal recording habits will also hinder standardization.

After analyzing the existing data, we found that the device name is generally composed of four parts: grid name, voltage level, device number, and device’s semantic name. The difficulty of equipment identification lies in the diversity of the expressions of each part and their combination.

(1) Diversity of grid names
The name of the grid is generally composed of place names, in which there is often a lack of superior place names, abbreviations of place names, and spaces.

(2) Diversity of voltage level expression
In the expression of voltage level, there is often a mixture of Chinese and English or missing expression in the unit.

(3) Diversity of device number expression
Some device numbers have the prefix "]#”, while some use numbers in Chinese, such as “四(four)”.
(4) Diversity of device’s semantic expression
Abbreviations often appear in this case. For example, “母线(busbar)” is recorded as “母” and “变电所(transformation substation)” is recorded “变”, etc.
(5) Diversity of combination
The difference in the order of parts and the way of dividing makes the device name different, such as “Tangshan. Balizhuang 35kV#4 B busbar”, “35kV #4 B busbar Balizhuang”, “4th B busbar Balizhuang 35kV”, etc.

The difficulty in identification of equipment names lies in the following aspects: firstly, the establishment and storage of professional lexicons. Power grid industry, as a specialized multidisciplinary intersection of physics and communication, its equipment is composed of professional vocabulary of the industry and combination of conventional pre-corpus, forming a comprehensive complex name. Secondly, due to the continuous text, absence of spaces in characters and numbers, different orders and ambiguity in segmentation, the identification of the device name is more complicated than processing pure Chinese or English names. In addition, due to the accumulation of historical data, the long-term lack of geographic and address name standardization management in the system, and the multiple expressions, abbreviations and records used because of staff personal habits also increase the difficulty of device name recognition.

3. Power Device Name Recognition Based on Tri-fold Search Tree
For the diversity of the various components and their combinations, this paper uses the dictionary database to analyze the full path name of the device, and recognizes the similarity between each part and the target device name. Semantic similarity and word order similarity are considered comprehensively in the similarity calculation.

3.1. Power device name dictionary construction based on tri-fold search tree
Compared with the other parts of the full path name of the grid equipment, the expression of the voltage level is relatively standardized. In order to save the storage space of the dictionary and improve query efficiency, we only include the place name, number and device semantic name in the dictionary construction. At the same time, these three parts are uniformly stored in the tri-fold dictionary search tree to improve query efficiency and the ease of maintenance.

The tri-fold dictionary search tree is a special Trie tree data structure, with each node used to store a character so as to complete the tree using the entire ordered data set. The tri-fold dictionary search tree has the efficiency advantage of digital search tree and the spatial advantage of the binary search tree, and is suitable for the construction and query of professional vocabulary [5].

The tri-fold dictionary search tree takes the intermediate value as the root node. The parent node forms a word with the direct child node. As for the left and right ones, it forms the first character of the adjacent word but not a word. The complexity of its query is log_c n. The tree building algorithm is shown in Figure 1.
Figure 1 Building algorithm of tri-fold dictionary search tree

3.2. Word segmentation of full path name of device

3.2.1. Digital number regularization
In the current local power systems, the device name recording habits vary from person to person, and the different representations of the same device are very common. Therefore, before regularizing the original device name, in addition to the preprocessing of encoding processing and non-Chinese character filtering, the unification of digital writing is also a very important operation. For digital records, there is no uniform norm in all walks of life. For example, the “Mazhuang 220kV#4 busbar” has the following expressions: “Mazhuang 220kV4 busbar”; “Mazhuang 220kV#4 busbar”; “Mazhuang 220kV 4th busbar”. This article uses a regular approach to regularize number matching in device names so as to convert numbers into uniform Arabic numbers.

3.2.2. Word segmentation
The device name describes the implementation of the device definition and is a given tag for a specific device so as to distinguish it from other devices. The staff can understand the name of the device and extract the information is because of the knowledge and habitual description of the device. Computers need to learn the knowledge related to name resolution in order to understand the semantic information in the device name. Since the input is a string, and the computer does not have an overall understanding of the entire string, Chinese word segmentation which converts the device name string into a component word array is needed, so that the computer can sequentially identify the words of the device string.

Common Chinese word segmentation methods include string matching based methods, understanding based methods, and statistical machine learning based methods. The word segmentation based word segmentation algorithm is fast in speed, simple in implementation and low in time complexity (O(n)), so this paper adopts this method as the basic word segmentation strategy.

The word segmentation algorithm based on string matching (scanning string) works as follows: the Chinese character string to be analyzed is matched with the term in a “sufficiently large” machine dictionary using a certain strategy. If a character string is found in the dictionary, then the match is successful (a word is recognized).

Based on scanning directions, the string matching word segmentation method can be divided into forward matching and reverse matching. Based on length priority matching, it can be divided into the largest (longest) matching and the smallest (shortest) matching. Based on whether or not to be combined with the process of marking part of speech, it can be divided into simple word segmentation method and an integrated method combining word segmentation and part-of-speech tagging.

The word segmentation often has problems with ambiguity and unregistered word segmentation. For example, the term “photovoltaic power station” in “Tengyuan Huichuan Photovoltaic Power Station 220kV#5 Busbar” may have never appeared, so the word segmentation granularity is an important indicator to be considered. This paper circumvents such problems according to the results of
the maximum matching word segmentation. The completed word segmentation algorithm is shown in Figure 2.

### Algorithm 2 Chinese words segmentation

| Input: | electrical equipment names |
|--------|----------------------------|
| Output: | equipment names segmentation array |

1. calculate the character length of the input string
2. record the match start location
3. while the record start location is less than the character length do
4. record the word with the maximum forward length
5. if this word matches a word in the thesaurus then
6. output the word and move the pointer down one bit
7. else
8. divide it into individual words, output single words and make the pointer point to the next bit
9. return result

**Figure 2 - Chinese word segmentation algorithm**

### 3.3. Semantic similarity matching

In the current dispatching system of the power grid, different representations of the same device are very common. For example, the “Mazhuang 220kV#4 busbar”, has several common expressions: “Mazhuang 220kV busbar”, “Mazhuang 220kV#4 busbar”, and “Mazhuang 220kV 4th busbar”. To identify whether these expressions describe the same device, it is necessary to design an appropriate similarity calculation method based on word segmentation. Unlike traditional natural language processed document-word vector space models, device names usually appear in a sentence, so its similarity should be measured within the sentence. Similarity calculation should consider not only the semantic similarity between normalized words, but also the word order and the similarity between Arabic numerals and special characters. Therefore, this paper uses a similarity calculation algorithm based on semantics and word order.

Suppose the set of word breakers for a given two devices is, \( T_i = \{ w_{i1}, ..., w_{im_i} \} (i \in \{1, 2\}) \), the constructed set is \( T = T_1 \cup T_2 = \{ w_1, ..., w_n \} \). Let \( S_i \) and \( R_i \) be the semantic expression vector and word order expression vector of two devices for \( T \) respectively. The similarities are \( S_S \) and \( S_R \) respectively, then the final similarity is weighted sum of the two.

\[
S = \alpha S_S + (1 - \alpha) S_R
\]

The semantic expression vector is a Boolean vector, that is, if a corresponding word is included in \( T \), then the corresponding component is 1, otherwise 0. The component of the word order vector corresponds to the sequence number of the word in \( T \) in the device word segmentation; if it does not appear in the sequence number set to its synonym or the given default value, then the algorithm is shown in Figure 3.

### Algorithm 3 calculate similarity

| Input: | equipment word segmentation \( T_1 = \{ w_{11}, w_{12}, ..., w_{1m_1} \}; \ T_2 = \{ w_{21}, w_{22}, ..., w_{2m_2} \} \) |
|--------|------------------------------------------------------|
| Output: | similarity \( S \) |

1. combine \( T_1 \) and \( T_2 \), remove redundant, denote as \( T = \{ w_1, w_2, ..., w_n \} \)
2. equalize and replace special characters in vectors \( T \)
3. for \( k = \{1, 2\} \) do
4. for \( i = \{1, 2, ..., n\} \) do
5. if \( w_i \) appears in \( T_k \) then
6. \( S_S = \{w_i \rightarrow 1\}, R_k = \{\text{Sequence in } T_k\} \)
7. else
8. \( S_S = \{w_i = 0.2\}, R_k = \{w_i = 0.4\} \)
9. end if
10. end for
11. end for
12. calculate semantic similarity \( S_S = \frac{s_S \cdot S_R}{\|S_S\| \cdot \|S_R\|} \)
13. calculate word order similarity \( S_R = 1 - \|\frac{S_S - S_R}{\|S_S\|}\| \)
14. calculate similarity \( S = \alpha \cdot S_S + (1 - \alpha) \cdot S_R \)
15. return \( S \)

**Figure 3 - Similarity calculation**
4. Design of Experiment and Result Analysis
This article uses JDK 1.7.0 as the development platform, PostgreSQL 9.4.5 as the database platform, and develops on the Windows 10 operating system.

4.1. Dataset construction
In order to verify the method proposed in this paper, 3270 pieces of log is extracted from the real equipment names from different business departments of the power system as experimental data. Based on this, two data sets are constructed for word segmentation experiments and similarity matching experiments: 1) Randomly extract 450 pieces as the data set for word segmentation experiments; 2) Randomly extract 100 pieces as positive samples, and manually simulate 100 pieces as negative samples. Match the manually labeled correct word segmentation result and similarity matching result to evaluate the experiment.

4.2. Evaluation indicators
This paper evaluates the word segmentation results by referring to the commonly used evaluation criteria of Chinese word segmentation, namely, precision, recall and error rate. The formula is as follows:

\[
\text{Precision} = \frac{TP}{TP + FN} \quad \text{Recall} = \frac{TP}{TP + FN} \quad \text{Error Rate} = \frac{FN}{N}
\]

TP is the number of words that are correctly divided; FN is the number of words that are incorrectly divided; N is the number of words divided based on gold standard. Among them, TP and FN are the number of differences in segmented words and the number of words correctly and incorrectly labeled.

4.3. Chinese word segmentation
In order to compare with word segregation using the tri-fold search tree dictionary in this paper, we construct a sequential dictionary based on binary tree search. The result of the segmentation comparison experiment are shown in Table 1.

As can be seen from Table 1, the error rate of the word segmentation results using the three-chapter dictionary search tree is only 3.13%, and the precision and recall rate are greatly improved.

| Condition of word segregation | Precision | Recall | Error Rate |
|------------------------------|-----------|--------|------------|
| Search tree dictionary       | 96.82%    | 95.33% | 3.3%       |
| Ordinary dictionary          | 76.86%    | 90.11% | 27.13%     |

4.4. Similarity matching
Since the result obtained by the device name similarity is a specific decimal, the threshold value of 0.88 is set as the criterion for judging the false positive according to the experimental result, that is, the predicted negative sample similarity > 0.88 is the predicted matching. The results of the confusion matrix of the device name similarity matching experiment are shown in Table 2. It can be seen that the positive sample is well predicted, which also proves the effect of the lexicon establishment and word segmentation. On the other hand, the threshold value of 0.88 leads to 44% positive example out of negative samples. Because of the professional power plant business scenario, the experiment requires the top N equipment with higher similarity, professional manual judgment of the staff to ensure the work is carefully performed. The threshold is set in the experiment for this purpose as well.
5. Engineering Application of the Algorithm

While constructing the core algorithm, this paper describes in detail how the algorithm can be applied to engineering. The scenario described is: if one or more newly generated maintenance texts are known, extract the power equipment information contained. Directly find the type of the corresponding device and similar IDs of the multiple data tables, and display the names of these similar devices according to the similarity order, and then conduct simple manually selection for manual or automatic registration.

The entire engineering application process is shown in Figure 4.

![Figure 4 Engineering Application Flowchart of Power Equipment Identification](image)

5.1. Pretreatment

In order to improve query efficiency and reduce the number of querying irrelevant databases, it is necessary to classify and manage the dictionary. A power device name string can be divided into place name, voltage level, number, and device. Therefore, a classification device table should be established first, and a dictionary data structure should be preset in the system to correspond to the name of the search table. Then, complete the voltage level in each device term and the key value of the locality. The number information can be expanded with public lexicon but not the data column.

If the number of dictionary entries is too large when the system starts, then load the dictionary tree into the memory in advance to save the time spent after the word segmentation.

5.2. Maintenance information processing

Take a maintenance text as an example: “Jibe. Gaositai/220kV.(#)4 busbar transfer repair; (Jibe) Gaositai (220kV) 2213 switch transfer repair”, in which what is in the parentheses is the normalized naming information omitted when filling out the repair text. First, the word segmentation operation can be divided into 10 words: “Jibe”, “Gaositai”, “220kV”, “4”, “busbar”, “transfer maintenance”, “Gaositai”, “2213”, “switch”, and “transfer repair”. The equipment for maintenance is from the two data tables of the busbar and the transformer, and the original string is divided into two substring lists according to the device name, that is, [“Jibe”, “Gaositai” , “220kV”, “4”, “busbar”] and [“Gaositai”, “2213”, “switch”].

5.3. Data screening and similarity calculation

After segmenting and extracting the keywords, the device keyword “busbar”, voltage level keyword “220kV”, and first-level place name keyword “Jibe” and second-level place name keyword “Gaositai” are queried for the first string list, so “Jibe. Gaositai/220kV.#4 Busbar”, “Jibe. Gaositai/220kV.#5 Busbar” can be directly queried in the table. After word segmentation, conduct similarity calculation with [“Gaositai”, “2213”, “switch”], and obtain a list of data structures
including the device ID, the normalized name, and the similarity with the comparison string. Return after sorting according to the similarity degree.

The second string list is queried with the first-level place name and voltage level ignored since they are missing. Then perform similarity calculation with [“Jibeil”, “Gaositai”, “220kV”, “4”, “Busbar”], and obtain a list of data structures including the device ID, the normalized device name, and the similarity with the comparison string. Return after sorting according to the similarity degree.

5.4. Performance Evaluation

In order to obtain the performance gap between the tri-fold dictionary tree segmentation and the ordinary linear dictionary segmentation, we conducted a control variable experiment with different data quantities, using the tri-fold dictionary tree segmentation and the ordinary linear dictionary segmentation used in this paper to perform word segmentation operations, and record the running time of each word segmentation, as shown in Table 3.

| Number of entries | 1024 | 2048 | 4096 |
|-------------------|------|------|------|
| Time of tri-fold word segmentation (ms) | 88   | 110  | 167  |
| Time of ordinary word segmentation (ms) | 1885 | 3329 | 6090 |

5.5. Post-processing

After the device similarity ranking table is obtained, the device involved in the maintenance text can be identified by simple manual identification, thereby performing registration or other processing methods. When developers use code, they can clean up large dictionary trees in memory as needed.

6. Conclusion

Deep exploration of equipment operation and maintenance data in the grid system can provide intelligent decision support for scientific assessment, accurate warning and targeted maintenance of equipment defects and faults. However, due to the heterogeneity of the operation and maintenance information systems at all levels of the power grid and difference of automation degrees, the device names are expressed differently in different systems, seriously hindering the data utilization efficiency.

Based on the in-depth analysis of the expression characteristics of power equipment names in various operation and maintenance systems of power grids, this paper constructs a device name dictionary database based on the tri-fold dictionary search tree and develop a device name word segmentation algorithm and a semantic and word order similarity calculation method to identify the target device. The next step is to enrich and optimize the structure and content of the dictionary tree to improve query accuracy and efficiency, and to explore equipment operation and maintenance data to achieve scientific evaluation and accurate warning of the equipment.

Acknowledgment

The science and technology project of STATE GRID Corporation of China - Research and Application of Event-handling Technology for Grid Monitoring Business Based on Machine Learning (SGJSNJ00FCJS1800810)

References
[1] Xi W, Xiaokai Z, Binyu X. Intelligent Substation Fault Diagnosis Method Based on Data Mining Technology[J]. Smart Power, 2018.
[2] Feifei Z, Ruitie F, Guanghui S, et al. Study on Substation Intelligent Monitoring Scheme Under Big Data Environment[J]. Shaanxi Electric Power, 2016.
[3] Junfeng Z, Zheyn Z, Yiqing C, et al. Intelligent power grid model under visual angle of RDF Semantic Web[J]. Huadian Technology, 2014, 36(4):19-21.
[4] Yan Y, Sheng G, Qiu R C, et al. Big Data Modeling and Analysis for Power Transmission Equipment: A Novel Random Matrix Theoretical Approach[J]. IEEE Access, 2017:1-1.
[5] Kamath C, Fan Y J. Using data mining to enable integration of wind resources on the power grid[J]. Statistical Analysis and Data Mining, 2012, 5(5):410 - 427.
[6] Varaiya P P, Wu F F, Bialek J W. Smart Operation of Smart Grid: Risk-Limiting Dispatch[J]. Proceedings of the IEEE, 2011, 99(1):40-57.
[7] Yao Y, Huang Z. Bi-directional LSTM Recurrent Neural Network for Chinese Word Segmentation[J]. 2016.
[8] Nan-Nan X U, Dong-Feng W, Pu H. Recognition of Two-Ticket System in Power Station Based on Forward Maximum Matching Algorithm[J]. Computer Simulation, 2014.
[9] Zhibin X, Jianfeng Z. FORWARD MAXIMUM MATCHING ALGORITHM BASED ON IMPROVED TRIE TREE STRUCTURE[J]. Computer Applications & Software, 2014.
[10] Deng D, Li G, Feng J, et al. Top-k string similarity search with edit-distance constraints[C]// 2013 IEEE 29th International Conference on Data Engineering (ICDE). IEEE, 2013.
[11] Soori H, Prilepok M, Platos J, et al. Utilizing Text Similarity Measurement for Data Compression to Detect Plagiarism in Czech[C]// Afro-european Conference for Industrial Advancement. 2015.