2021 IEEE International Conference on Electrical Engineering and Mechatronics Technology (ICEEMT) | 978-1-6654-4128-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICEEMT52412.2021.9602260

A Vertical Semantic Search Engine In Electric Power Metering Domain

1st LI Sheng
Metrology Center
Digital Grid Research Institute,
China Southern Power Grid
Guangzhou, China
lisheng@csg.cn

2nd ZHENG Kaihong
Metrology Center
Digital Grid Research Institute,
China Southern Power Grid
Guangzhou, China
zhengkh@csg.cn

3rd YANG Jinfeng
Marketing Department
China Southern Power Grid
Guangzhou, China
yangjingfeng@csg.cn

4th WANG Xin*
Zhejiang University-China Southern Power Grid Joint Research Centre on AI
Zhejiang University
Hangzhou, China
wangxin2009@zju.edu.cn

5th ZENG Lukun
Metrology Center
Digital Grid Research Institute,
China Southern Power Grid
Guangzhou, China
zenglk@csg.cn

6th GONG Qihang
Metrology Center
Digital Grid Research Institute,
China Southern Power Grid
Guangzhou, China
gongqh@csg.cn

7th YANG Geng
Metrology Center
Digital Grid Research Institute,
China Southern Power Grid
Guangzhou, China
yanggeng@cx.yn.csg.cn

8th YU Zhixi
School of Computer Science and Technology
Zhejiang University of Technology
Hangzhou, China
381669225@qq.com

9th LIANG Yongjie
School of Computer Science and Technology
Zhejiang University of Technology
Hangzhou, China
cs_lyj@163.com

***Abstract*—With the explosive growth of data on the Internet, the mining of vertical knowledge domain material has become more complex. In order to efficiently collect, manage and exploit the huge knowledge corpus, this paper proposes a vertical semantic search engine for the domain of electric power metering. On the one hand the engine overcomes the shortcomings of traditional general-purpose search engines in terms of lack of targeting and specialization, and on the other hand it has higher accuracy and more stable recall than general keyword-based search engines, enabling semantic understanding and relational analysis. The main innovation of the proposed engine is the use of knowledge units extracted from knowledge graphs for semantic search. Experimental results prove that this engine can improve the accuracy and timeliness of power metering knowledge retrieval, and has certain application value.**

Keywords-electric power metering; vertical search engine; knowledge graph; semantic search; knowledge unit

I. INTRODUCTION

In recent years, knowledge graph and search technologies based on knowledge graph have been widely studied and applied in the electric power domain. Gao H.X[1] et al. summarized the specific application scenarios of knowledge graph in the electric power domain and pointed out possible research directions of knowledge graph in the electric power domain. Knowledge graph connects a large amount of different kinds of information together to get a relationship network, which can express the relationship between entities intuitively and effectively. Large amounts of data resources including electricity consumption, relevant standards, technical documents, etc. in the electric power domain continue to accumulate. The application of knowledge graph can greatly improve the efficiency and accuracy of information search in the electric power domain.

There are many applications of knowledge graph in the electric power domain Representative works are follows: Jiang Wei et al.[2] constructed the infrastructure engineering knowledge graph based Neo4j and realized intelligent retrieval function based on the graph. Tang Y.C et al. [3] presented and developed an integrated management system of power equipment quality based on graph database and knowledge graph for the problems in power equipment management, and a lot of problems have been effectively solved by this system. Wu Chao[4] combined Transformer, Bi-GRU and Attention models to design a knowledge extraction system in knowledge graph of electric power dispatch. The system realized the automatic knowledge extraction in the early stage of the construction of knowledge graphs in the power domain, which reduces costs and improves work efficiency. Based on the characteristics of distribution network knowledge graph and distribution dispatch text, Zheng W.Y et al. [5] proposed an entity linking method of distribution dispatching texts for a distribution network knowledge graph. This model structure of a Lexical Semantic Feature-based Skip Convolution Neural Network (LSF-SCNN) was improved. And the experimental results show that the method obtains a high accuracy rate in

978-1-6654-4128-5/21/$31.00 ©2021 IEEE

entity linking. Tan Gang et al. [6] constructed a hybrid domain feature knowledge graph smart Question Answering system (HDKG-QA), which reduced the fuzzy of Chinese language questions and the high cost of online service operation and maintenance in a KG based QA. Wang Yuan et al. [7] proposed a knowledge graph construction method based on the full-service unified data center. Experiments show that these method improves the precision and recall rate, and has better intelligent search and analysis ability.

In terms of semantic search research in the electric power domain. Qiu Jian et al. [8] took the circuit breaker as a case in point to establish a framework of text mining-based life-cycle condition assessment. Ji Yuan et al. [9] designed a construction method of semantic search system in power domain. The method used intelligent domain word segmentation technology to extract semantic knowledge, realized semantic based on knowledge graph, and achieved great search results. Guo Y.Y[10] designed a knowledge graph-based grid search engine using the Neo4j graph database to greatly improve the efficiency of grid workers in accessing information.

In addition to the electric power domain, Wang Hong et al. [11] proposed a domain ontology-based semantic retrieval framework, and the model can solve the problems of existing search techniques in the information retrieval of civil aviation airport emergency resources. This research also has certain enlightening significance for the work of this paper.

This paper designs a semantic search engine which aims to solve the data retrieval problem in the domain of power metering. Power metering data including electricity supply, electricity consumption, etc. And building a knowledge graph in the domain of power metering can greatly improve the efficiency and accuracy of search engines for power metering data retrieval.

II. CONSTRUCTION OF A VERTICAL SEMANTIC SEARCH ENGINE

A. Information collection

In terms of information collection, our system has built a crawler system architecture based on Deep Web crawler in response to the wide distribution of electricity data and the diversity of data structures. The system can first obtain indexed lists (URLs) through traditional search engines, and then access the Web pages hidden behind the basic static pages by submitting the corresponding keywords or performing certain interactions, thereby improving the system's knowledge mining capability.

In order to improve the relevance between the crawled knowledge resources and topics, our system uses link relevance analysis techniques and topic relevance analysis techniques.

1) Link relevance analysis techniques

Link relevance analysis is a coarse filtering of links to be crawled from a web page, implemented before the text crawling effort. This step is used to determine whether the initially obtained links (URLs) to a web page have substantial content, or to determine in general whether the page provides a large number of links to other electricity metering related pages.

The system uses the Hits algorithm for link relevance analysis. The calculations are shown below.

There are three pages A, B and C and the Authority and Hub of each page are 1. The relationship is shown in Figure 1.

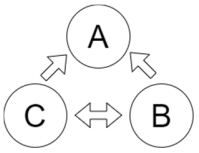

Figure 1. Relationship between pages A, B and C

Let the value of authority be ay and the value of hub be hb, and after performing one iteration we get ay(A)=2, ay(B)=1, ay(C)=1; hb(A)=0, hb(B)=2, hb(C)=2.

Therefore, for the collected web links (parent links), the authority and hub values need to be analyzed firstly. Only if the two indicators exceed a certain threshold, the page is saved and the child links are extracted from the page and placed in the queue to be crawled; otherwise the parent links will be discarded.

2) Topic relevance analysis techniques

Topic relevance analysis is the process of selecting textual information on a page to perform a similarity calculation with the topic of electricity metering, in order to judge the relevance of the page to the topic.

A simple vector distance algorithm based on a vector space model is used in the system. The basic idea is to represent the text information in a web page as an N-dimensional vector, and use the weights of the feature terms as components, and use the size of the angle between the N-dimensional vector and the components to represent the similarity of things. The smaller the angle between the vectors, the more similar the thing represented by the two vectors is.

B. Information extraction

Information extraction is the process of crawling a web page for text information and storing it in memory. Before the information is formally extracted from the web page, some filtering of the crawled content is carried out.

An HTML page can be divided into several areas, such as the top bar, sidebar, main area, bottom bar, etc. The text information displayed in different areas conveys different information usefulness. At the same time, each area is divided into smaller sections with different tags (e.g. <h1>, , <title>, etc.), and these sections differ in their degree of relevance to the topic.

Therefore, an HTML page can be considered as a DOM (Document Object Model) tree, with different areas of the page as branches of the tree and the small chunks represented by different tags as nodes of the tree, and the STU-DOM model is used to calculate the relevance of each node to the topic.

If the topic relevance of the chunk is greater than Y, the chunk is considered to be more relevant to the topic and the chunk is crawled; if not, the chunk will be discarded.

C. Information Retrieval

A document usually contains many terms. Each of which has a different relevance to the document topic. So it is necessary to determine the document feature items by calculating the weight of each term before retrieval. The document relevance calculation consists of two parts: the calculation of the weight of the terms in the document and the calculation of the relationships between the terms. Here, only the weight of the terms in the document is used, and the words with the highest weight are defined as feature items. The calculation formula is

$$W_{t,D}=tf_{t,D}\times \log(n/Df_t)$$

t means the term and D means document. $tf_{t,D}$ indicates the number of times a term is mentioned in a document. The more times the term is mentioned in the same document, the more relevant the word is to the subject of the document. Df_t indicates how many documents contain the participle. When a high number of documents containing the term it indicates that the term is likely to be generic and not meaningful.

Adjust the weight of the term according to where it appears in the document, keywords, abstracts, citations, etc.

As search statements are generally long, it is a prerequisite for the search function to be able to break them down and extract the keywords from them.

For Chinese word segmentation, the system uses a rule-based and dictionary-based Chinese word segmentation algorithm. In other words, the system uses a forward iterative fine-grained segmentation algorithm to iteractively segment the search statement from the largest word to the smallest word, thus turning long sentences into short words. The remaining words are used as search keywords after eliminating words without real meaning, such as intonation words and auxiliary verbs.

III. Semantic search for power metering

A. Knowledge unit extraction and entity creation

The extraction of knowledge units is the core of the establishment of the semantic search function. As the knowledge units are stored in the knowledge graph as a triple, it is necessary to distinguish between entities and relations and to correspond them when extracting them. The process of extracting knowledge units is illustrated in Figure 2.

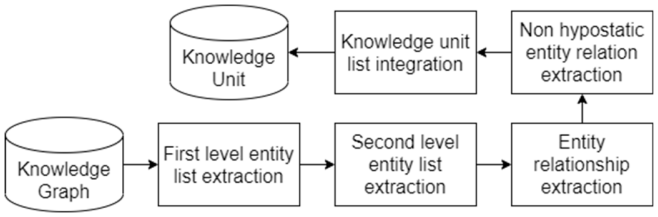

Figure 2. Knowledge unit extraction process

Unlike the usual way of storing textual data, knowledge graphs are often stored using graph databases, so that a tree-like structure can be clearly represented. Starting from the root node, knowledge units are extracted from the top down, and the knowledge points with a large number and covering a wide range of knowledge are classified according to the distance of each child node from the root node. The distance of a child node from the root node is judged using the shortest path. Unit distance is represented by one edge. If the path from the child node to the root node passes through at least two intermediate nodes, then the path from the child node to the root node is four units.

Our system divides the entities into 2 levels, and first extracts the first level entities. Let the depth of the root node be 0, then all nodes with depth 1 (i.e. children directly connected to the root node) are treated as level 1 entities. At this point, all nodes with depth 2 are taken as level 2 entities because level 1 entities cover a wide range of knowledge and can also be subdivided into smaller entities.

B. Semantic search based on knowledge units

Semantic search means that the scope of the search engine is not limited to the search statement itself, but to dig deeper into the meaning of the words, that is, through the phenomenon to see the essence. For example, if you enter the search term "substation", the results will not only contain the word "substation" directly, but also show the content related to the power station.

The knowledge units extracted in this paper are stored within a MySQL database. The database management system uses SQL language to realize semantic search.

SELECT b FROM a WHERE a.b contains the ‘substation’ AND b is a data exception

Using the simple query statement above, substations with abnormal data problems can be found and the type of abnormality is also displayed. Compared with ordinary keyword search, semantic search broadens the scope of search, improves the versatility of query results, and can find more query sentence related data as much as possible.

This paper uses the MaLSTM [12] model to implement a semantic metric that is based on Manhattan distance, which is superior to the original cosine degree similarity and Euclidean distance. The model takes distance as the goal, uses LSTM (long-term and short-term memory network) to model the complex semantics, obtains two semantics with fixed length, and then uses the following formula to calculate the similarity.

$$g\left(h_{T_m}^{(m)},h_{T_n}^{(n)}\right)=\exp\left(-\left\|h_{T_m}^{(m)}-h_{T_n}^{(n)}\right\|_1\right)$$

where h_t denotes the hidden layer state, g is the similarity of semantics m to n ,$g\in(0,1)$.

IV. Experiments and analysis

To evaluate the efficiency of a vertical semantic search engine for the domain of power metering, system tests were conducted within the Southern Power Grid Company. The test data collected included a total of over 40,000 pieces of information on grid plant stations, grid lines, grid staff and circuit fault records, a total of over 2 million electricity consumption records for three months, and 900 pieces of related literature materials. The web spider crawled to over 100,000 pieces of data, with data sources including general

search engines, industry-related websites, literature search websites, etc.

When a web spider is crawling, the data is stored in two ways. One is to construct knowledge base forms and store them with direct reference to the data obtained by the crawl; the other is to split and store the data based on the data table fields obtained by analyzing the knowledge units. A keyword search is performed on the former form and a semantic search is performed on the latter form to achieve a typical lookup task, and the results are compared and analyzed. The experiments demonstrate that semantic search outperforms keyword search by accuracy and recall, where accuracy refers to the percentage of correct results in the returned results and recall refers to the ratio of correct results to the actual existence of correct results. The larger the value of accuracy and recall, the better the search effect. As shown in Table 1, seven common search criteria were selected and SQL queries were performed using Like statements.

TABLE I. TABLE OF SEARCH CRITERIA

No.	Search criteria
1	A moment of sudden change in Guangzhou's electricity data
2	Number of faults in feedback data from a substation
3	The moment when electricity data peaks in a region
4	Substation locations not maintained for three months
5	Engineers with extensive maintenance experience
6	Towns that exceed the regional electricity load
7	Number of substations in Guangzhou that suffered a lightning fault within 15 days

Figures 3 show the experimental results corresponding to the seven search criteria.

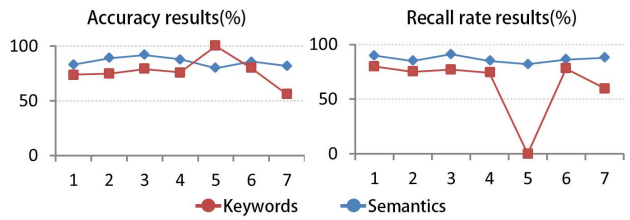

Figure 3. Accuracy results and recall rate results

As can be seen from the figures, the query results obtained using semantic search generally outperformed the keyword search results in terms of accuracy and recall. The reason that the two values fluctuate less when using semantic search is that semantic search also takes into account the inter-entity relationships compared to keyword search, which only takes into account the character string matching. For example, for search criteria 5, the keyword search has an accuracy of 100% but a recall rate of 0% due to the inability to understand the meaning of "extensive maintenance experience", whereas the semantic search can analyse the relationship between experience and working years. For search criteria 7, the difference in accuracy and recall rate is greater because the keyword search can only split the words "lightning strike" and "fault" and then search for them separately, whereas the semantic search can find the relationship between the two words, highlighting the superiority of the semantic search.

V. CONCLUSION

In this paper, we propose a vertical semantic search engine for power metering to meet the needs of users in finding power metering knowledge, and present the whole process of knowledge collection, knowledge storage, knowledge unit extraction and semantic search. On the one hand, it proves the effectiveness and application value of our method in the processing of big data of power metering through platform testing and a large amount of real data; on the other hand, the vertical semantic search engine has significantly improved the accuracy and efficiency of search results compared with existing search engines.

However, our work is still in the early stages of research on the diversity of applications of knowledge units in the domain of power metering. How to improve the efficiency of semantic search still needs to be further explored by combining deep learning, natural language processing and other techniques.

ACKNOWLEDGMENT

This research is supported by the Science and Technology Project of China Southern Power Grid Co., Ltd. (ZBKJXM20190046) and the National Key R&D Program of China (2020YFB0906004).

REFERENCES

[1] GAO Haixiang, MIAO Lu, LIU Jianing, LIN Xiangning, DONG Kai, HE Xiangzhen. Review on knowledge graph and its application in power systems [J]. Guangdong Electric Power,2020,33(09):66-76.

[2] JIANG Wei, ZHOU Ying, CHENG Shuqi, WANG Bo, CHENG Hong. Research on the power grid project data mining and knowledge graph construction technologies [J]. Electric Power ICT, 2021,19(02):15-22.

[3] TANG Yachen, FANG Dingjiang, HAN Haiyun, et al. Research on power equipment quality integrated management system based on graph database and knowledge graph [J]. Distribution & Utilization, 2019, 36 (11): 35-40.

[4] WU Chao. Design and implementation of knowledge extraction system in knowledge graph of electric power dispatch[D]. University of Chinese Academy of Sciences (Shenyang Institute of Computing Technology, Chinese Academy of Sciences),2020.

[5] ZHENG Weiyan, YANG Yong, LU Jiaju, ZHENG Jie, TAN Haiyun, YU Jianfei, YU Tengfei. Entity linking method of distribution dispatching texts for a distribution network knowledge graph [J]. Power System Protection and Control,2021,49(04):111-117.

[6] TAN Gang, CHEN Yu, PENG Yunzhu. Hybrid domain feature knowledge graph smart question answering system. Computer Engineering and Applications, 2020, 56 (3): 232-239.

[7] WANG Yuan, PENG Chenhui, WANG Zhiqiang, et al. Application of knowledge graph in full-service unified data center of national grid. Computer Engineering and Applications, 2019, 55 (15): 104-109.

[8] QIU Jian, WANG Huifang, YING Gaoliang, ZHANG Bo, ZOU Guoping, HE Benteng. Text mining technique and application of lifecycle condition assessment for circuit breaker[J]. Automation of Electric Power Systems, 2016,40(06):107-112+118.

[9] JI Yuan, XIE Dong, ZHOU Siming, WANG Xiangdong. Construction method of semantic search system in power domain[J]. Computer Systems Applications, 2016,25(04):91-96.

[10] GUO Yunying. Design and implementation of grid information search engine based on knowledge graph[D]. University of Chinese Academy

of Sciences (Shenyang Institute of Computing Technology, Chinese Academy of Sciences),2020.

[11] WANG Hong, XIAO Zhiwei, LI Jian, FAN Hongjie. Research on semantic retrieval for civil aviation airport emergency resource[J]. Computer Applications and Software,2014,31(01):73-76+153.

[12] Thyagarajan A. Siamese Recurrent Architectures for Learning Sentence Similarity [C]// Thirtieth Aaai Conference on Artificial Intelligence. AAAI Press, 2016.