Construction and Application of Knowledge Graph in Full-service Unified Data Center of Electric Power System

Yuan Wang[a], Keheng Zhang[b], Qicheng Dai[*], Chenhui Peng[c], Kang Zhao[d]

NARI Group Co. Ltd. , China Realtime Database Co. , Ltd. , Nanjing 210012, China.

*Corresponding author e-mail: daiqicheng@sgepri.sgcc.com.cn,
[a]wangyuan1@sgepri.sgcc.com.cn, [b]zhangkeheng@sgepri.sgcc.com.cn,
[c]pengchenhui@sgepri.sgcc.com.cn, [d]zhaokang@sgepri.sgcc.com.cn

Abstract. Aiming at the problems of insufficient coordination and sharing of cross-service information and the inadequate utilization of power big data assets, this paper proposes the idea of constructing a power system full-service data knowledge graph combining the background of constructing Full-service Unified Data Center of Electric Power System. Power big data resources can be managed and utilized more intelligently and efficiently by implementing semantic-level connectivity. The application architecture of the knowledge graph in the full-service unified data center is proposed, the key technologies of its construction are analyzed, and a prototype system is built. The system is deployed in the unified data center of State Grid Zhejiang and tested by two typical scenarios of intelligent search and blackout information analysis, the result shows that the knowledge graph can effectively improve the accuracy and recall rate of the search and realize the cross-service information collaboration, optimizing service process.

1. Introduction

Smart grid construction highly integrates information technology with power physics systems, thus making power data an explosive growth and becoming important asset for power grid companies [1]. Power data is generated in all aspects of power system, such as production and operation. It can fully reflect the operating status of the power system and the health status of the equipment. It also contains many physical mechanisms of the grid, complex event logic, expert knowledge and experience, etc., and has received extensive attention from academia and industry [2]. State Grid Corporation has realized the unified management of master data through the construction of the headquarters, provincial (city) two-level data center and the main data management platform. The amount of data accumulated has exceeded 5PB. However, at present, each service independently maintains a set of data systems according to its own needs, which causes insufficient coordination and sharing of cross-service information, duplication storage and read of data, poor real-time and reliable and other issues. In order to achieve data penetration of full-service processes, fully exploit and utilize the value of power big data, State Grid Corporation plans to construct a full-service unified data center with “clean and transparent data, unified and specific model, and flexible and intelligence analysis” at the end of the “13th five-year plan”, supporting unified storage, management and service of grid data. Studying the semantic network construction of power data, providing an intelligent and efficient power data retrieval method to support cross-professional service process optimization and ultilizing data assets is an important content of unified data platform construction [4].

Knowledge graph (KG) is a typical semantic network. Its concept was first proposed by Google in 2012 as the core of its intelligent search engine [5]. Traditional data retrieval technology directly establishes the index of the webpage then the contents are searched by keywords. KG extracts all the knowledge to form a graph structure, with nodes representing entities and edges representing relationships. Because the logical relationship between knowledge is included, it can understand user's intentions more accurately and return the correct answers. For example, when a user searches for "the population of Jiangsu Province", the system will return "8029 million people" rather than results containing the keyword " the population of Jiangsu Province ". In recent years, many IT companies around the world such as Microsoft, IBM, Baidu, Sogou, etc. have joined this field and launched their products. And because of its advantages in knowledge representation, the applications are no longer limited in search engine. Facebook and Apple Siri use it to scenarios of socialize and DeepQA to enhance the computer's understanding of user's behavior; financial companies use KG to correlate user information obtained from different channels to provide a basis for credit approval; KG can also be used to data matching in big data analysis.

KG has also been applied in the power industry. For example, in the customer service system, KG helps the customer representatives find information that the customer wants comprehensively, it can also improve the accuracy of the responses by intelligent customer service system [6]; Retrieval of equipment defects using KG to improve accuracy and recall [7]. State Grid has established a comprehensive knowledge base, a professional knowledge base and a thematic knowledge base, but has not yet established a knowledge graph covering all service data, to say nothing of implementing related applications. This paper presents the application architecture of the KG in the unified full-service data center. the key technologies of its construction are studied and a prototype system is built. The prototype system is deployed and tested in the unified data center of State Grid Zhejiang.

2. Architecture of Knowledge Graph in the Unified Full-service Data Center of Electric Power System

The knowledge graph, as shown in figure 1, is located between the unified storage area and the unified analysis area in the data center. The unified storage area stores all grid data in different databases according to the three data types: structured, semi-structured and unstructured. The unified analysis area extracts the required data to build the application model according to different application requirements. While traditional analysis service directly extract grid data, the knowledge graph additionally provides a data retrieval method which have considered including logic and semantics.

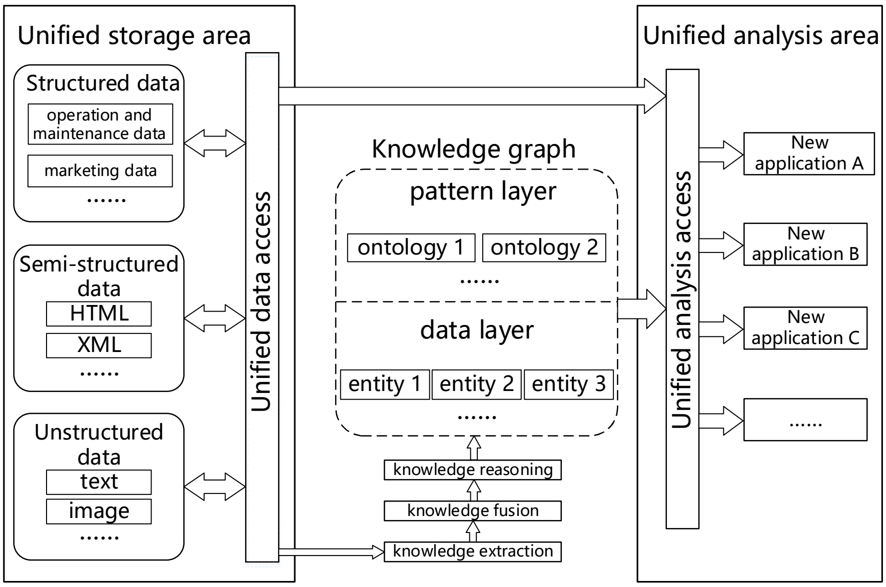

Figure 1. Architecture of knowledge graph in the unified full-service data center of Electric Power System

When receiving the paper, we assume that the corresponding authors grant us the copyright to use the paper for the book or journal in question. Should authors use tables or figures from other Publications, they must ask the corresponding publishers to grant them the right to publish this material in their paper.

Knowledge graph is essentially a knowledge base which uses a network to connect and organize entities and attributes [8]. It can be divided into two logical layers: the data layer and the pattern layer. The pattern layer which is the core of a knowledge graph is built above the data layer and regulates the factual representation of the data layer through the ontology library. The data layer stores the knowledge facts in the graph database, using the "entity-relation-entity" or "entity-attribute-value" triple as basic unit. Knowledge facts are formed from the process which includes extracting entities, relationships and attributes from data in the data center, eliminating errors and redundancy through knowledge fusion and improving structure through knowledge reasoning.

3. key technologies of constructing Knowledge Graph in the Unified Full-service Data of Power System

3.1. knowledge extraction

Knowledge extraction means extracting entities, attributes and relationships from data center and external database and constituting the basic element of the knowledge graph – the triplet. Most power data has been extracted and cleaned when stored in the data center, thus has became structured data.
These data are stored in relational databases with relatively explicit entities, attributes and relationships and can be directly converted to Resource description framework (RDF) [11]. The rules are shown in figure 2: 1) A table in a relational database becomes a class in RDF, for example, if there are 5 tables in mysql, then they become 5 classes in the converted ontology. 2) The column of the table becomes an attribute. 3) The row of the table becomes an instance. 4) The value of the cell in the table becomes a literal. 5) If the cell is a foreign key, then its value is an entity.

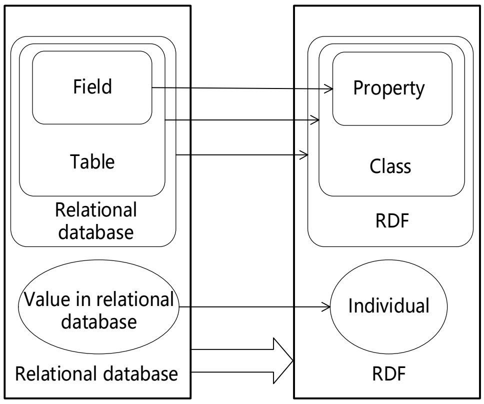

Figure 2. Relational database converting to RDF

There's also a significant amount of semi-structured and unstructured data such as XML, HTML, documents, images and audio/video. Text data of which the common extraction method is based on machine learning word segmentation or dictionary-based word segmentation is the main part of them. The segmentation method based on machine learning heavily relies on the training set, so it's difficult to achieve the desired effect in field word segmentation. Dictionary-based word segmentation methods have high accuracy, but it cost professionals a lot of time and effort. The power dictionary written in this paper which can achieve accurate word segmentation and matching is based on the comprehensive knowledge base, the professional knowledge base and the thematic knowledge base established by State Grid Corporation in the early stage. The process of knowledge extraction is as follows: 1) segment text of semi-structured/unstructured data using hidden Markov models. 2) Retrieve the words after segmentation one by one in the dictionary including the power professional ones and general ones, extract the word if there's a match. 3) Tag the part-of-speech to the word according to the dictionaries. Unstructured data like images and video streams occupies a relatively small part of all the power data, they need to be semantically labeled first, and then can be extracted same as the unstructured text data.

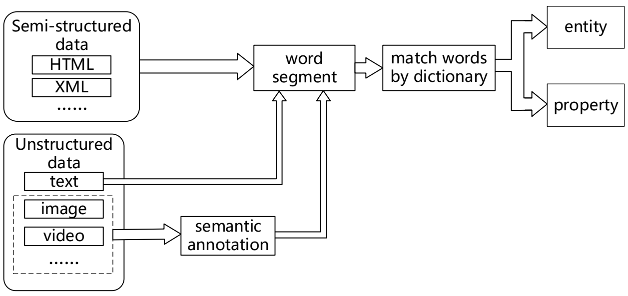

Figure 3. Extraction process of semi-structured and unstructured data

3.2. knowledge fusion
The purpose of knowledge fusion is to eliminate the ambiguity, error and redundancy of entities and relationships acquired by knowledge extraction and ensure the quality of knowledge. The structured database in the data center corresponds to the power knowledge base, so they have relative higher data quality, but there are also cases that a device owns different names, for example, “knife gate” and “isolation switch” refer to the isolating switch. "cutout switch" and "circuit breaker" refer to circuit breaker. Semi-structured data can't be directly converted to RDF because of their non-standard data formats, for example, different service systems use different description formats for the same device, so their entities and relationships need to merged into the structured database. Segmentation and

extraction of unstructured data based on the power dictionary for screening usually do not have a polysemy. However, some unstructured data like operation logs are manual record which may use different words to represent the same device. Therefore, the main work of knowledge fusion of grid data is to achieve coreference resolution. Firstly, the words can be divided into different groups according to the character. Pronouns are not used in grid service record, so synonym will be in the same group. Then calculate the cosine similarity between different words and judge their relationships. Some auxiliary rules can be considered for judging, such as synonyms usually do not appear in the same sentence and synonyms often appear at similar positions, etc. Finally, the words of the same meaning are formed into a set and a word is selected as a representative. Finally, the selected synonyms are formed into a set of words, and each word is selected as a representative.

3.3. knowledge reasoning

Entities and relationships are obtained from the original data after knowledge extraction and knowledge fusion. However, the ontology framework made by experts who are usually only familiar with their own field is difficult to cover all relationships, so the knowledge reasoning that is used to improve the existing knowledge framework and explore the potential logical relationship between different services is necessary. Graph-based reasoning method is adopted in this paper, entity in the graph represents node and relationship or attribute represent edge. If there's a certain path starts from node A and can reach node B, then we can speculate their relation by their semantic distance. The semantic distance between the two words w1 and w2 is the sum of the weights of the shortest path between the concepts they belong to:

$$Dist(C_1, C_2) = \sum_{i=1}^{n} weight_i \tag{1}$$

Where weighti is the weight of side i.

All edges from concept C have the same weight as weight(C):

$$weight(C) = \frac{1}{\alpha^{Dep(C)} * Wid(C)}, \alpha \geq 2 \tag{2}$$

Where Dep(C), Wid(C) are the depth and width of C, α is a variable parameter.

4. Application of knowledge graph in the unified full-service data center of Electric Power System

Knowledge graph provide a more effective management method for power data, making data organization more natural to the way of human thinking, and thus have unique advantages in the intelligent search and cross-service collaborative analysis. The prototype system is developed and deployed in the unified data center of State Grid Zhejiang, the function is test in two typical scenarios of intelligent search and blackout information analysis.

4.1. intelligent search

The intelligent search based on KG is actually a long tail search [14]. user's search content is semantically understood by the system through segmentation and matching, and linked to the corresponding knowledge in the base. Then the core concepts will be expanded to related ones that will be retrieved in turn. After comparing the similarity algorithm, the higher correlation results will be given. The similarity value is calculated according to formula (3).

$$sim(C_1,C_2)=\frac{1}{1+Dist(C_1,C_2)} \tag{3}$$

There are 1,000 questions frequently searched in typical scenarios in operation and maintenance system, dispatching system and material system tested with keyword search and knowledge graph-based search, the results are shown in Table 1. The accuracy and recall rate of KG is significantly improved compared with the search methods of keywords. For its advantage of accurately identifying keywords and achieving semantic understanding with strong pertinence based on the power knowledge base.

Table 1. Comparison of results from different retrieval methods

Method	Accuracy (%)	Recall (%)
Keyword	77.36	56.73
KG	90.19	91.27

In order to intuitively display the above advantages of KG, 6 typical queries about Xiangshan Substation and their results are listed in Table 2.

Table 2. Typical queries

Num	Query problem
1	Voltage level of Xiangshan substation
2	Operation state of Xiangshan substation
3	The outlet lines of Xiangshan substation
4	Operational personnel in Xiangshan substation
5	Equipment that has failed in Xiangshan substation in a year
6	Equipment that has failed more than twice in Xiangshan substation in a year

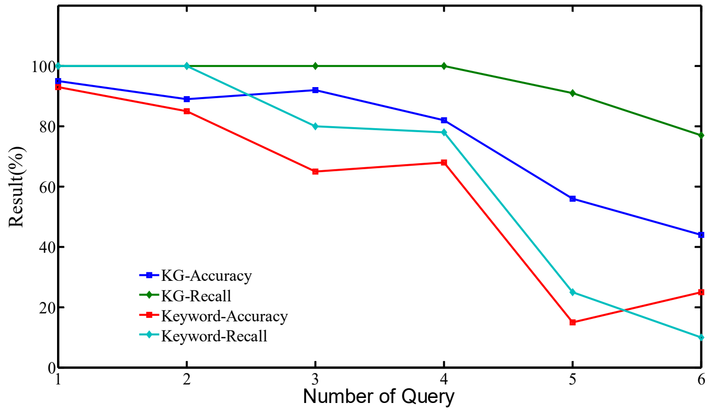

Figure 4. Results of typical queries

4.2. cross-service data collaboration – as an example of power outage plan information analysis

When the grid company formulates the power outage plan, realizing the automatic analysis and report of the power outage scope, equipment and user list can effectively improve the customer experience, the company's response capability and the lean management of the distribution network operation.

At present the process is complex and inefficient. The power supply service command platform at first compiles the power outage information according to the power outage plan. Then the dispatching system determines the power outage scope according to the plan, and finally the marketing system confirms the affected user list according to the power outage range.

The knowledge graph based on marketing, power distribution and dispatching system can link service data from different departments, clearly express the service process and achieve quick analysis of power outage information. First, establish the entities such as “substation”, “transformer”, “customer” and associate underlying data. Then construct relationships between the entities based on the knowledge base and save them in the graph database. Finally, it can be achieved based on the knowledge graph that the operator query power outage scope according to the initial power outage site, list the detailed user list, guarantee the power supply for important loads and active management for users of different levels.

As an example, the knowledge graph of a certain area in Zhejiang is constructed as shown in Fig 5. When the Qianchuan substation need to be cut out, the operator in power supply service command platform will see that A280 line and its customers will be affected and have to be blackout as well as 423 line, however the user Xie Jiankang can be supplied by Xixing substation. Then the power outage plan can be confirmed and the order will be given to dispatching and maintenance departments.

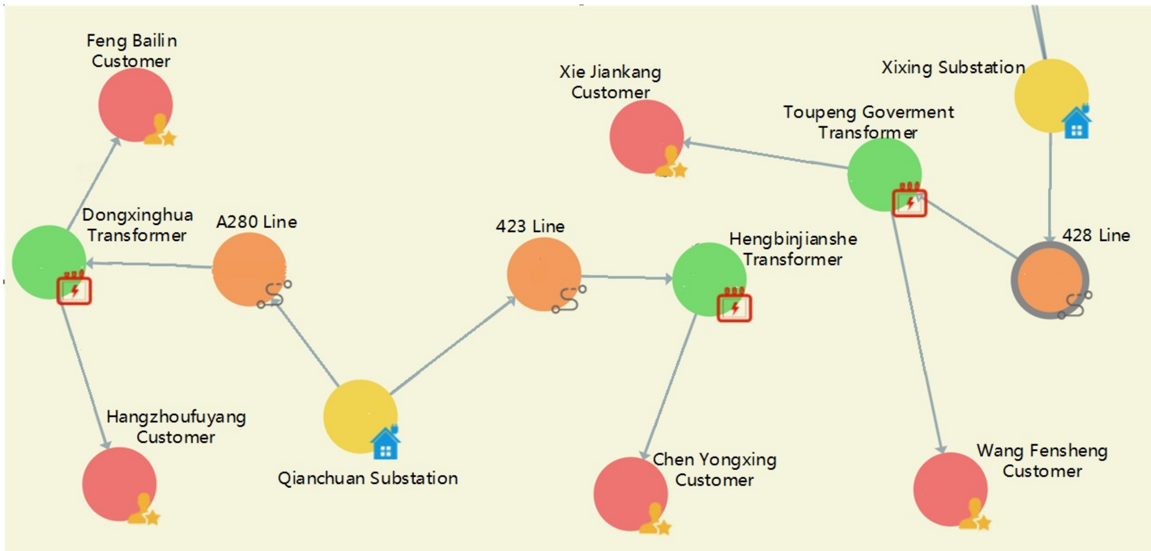

Figure 5. Knowledge graph of distribution grid in a certain area in Zhejiang

5. Conclusion

This paper presents a method of constructing knowledge graph in the unified full-service data center of electric power system which can organize and manage power big data smarter and more efficiently. The application architecture of the KG in the unified full-service data center is proposed, key technologies of constructing the KG are studied and analyzed. The prototype system is developed and deployed in the unified data center of State Grid Zhejiang. The application results of intelligent search and power outage information analysis show that the KG can effectively improve the accuracy and recall in electric power system, collaborate cross-service data and optimize service process.

Acknowledgments

This work was financially supported by State Grid science and technology project: Research on key technologies and application of unified data model of power grid based on Knowledge Graph (Project number: 5211XT180045) fund.

References

[1] Yu X, Xue Y. Smart Grids: A Cyber–Physical Systems Perspective [J]. Proceedings of the IEEE, 2016, 104 (5):1058-1070.

[2] Qiu Jian. Research on Chinese textual data mining techniques and reliability applications in power system. Hangzhou: Zhejiang University, 2016 (in Chinese).

[3] Zhu Biqin, Wu Fei, Luo Fucai. Research on the construction of data analysis domain of unified data center based on big data. Journal of Electric Power Information and Communication Technology, 2017 (2):91-96. (in Chinese).
[4] PENG Xiaosheng, DENG Diyuan, CHENG Shijie, et al. Key technologies of electric power big data and its application prospects in smart grid [J]. Proceeding of the CSEE,2015,35 (3):503-511. (in Chinese).
[5] AMIT S. Introducing the knowledge graph [R]. America: Official Blog of Google, 2012.
[6] Rao Zhuyi, Zhang Yunxiang. Research on intelligent customer service system based on knowledge map [J]. Journal of Electric Power Information and Communication Technology, 2017,15 (07):41-45. (in Chinese).
[7] Liu Ziquan, Wang Huifang. Retrieval method for defect records of power equipment based on knowledge graph technology [J]. Journal of Automation of Electric Power System, 2018,42 (14):158-164. (in Chinese).
[8] XU Zenglin, SHENG Yongpan, HE Lirong, et al. Review on knowledge graph techniques[J]. Journal of University of Electronic Science and Technology of China, 2016, 45 (4): 589-606(in Chinese).
[9] LIU Qiao, LI Yang, DUAN Hong, et al. Knowledge graph construction techniques [J]. Journal of Computer Research and Development, 2016, 53 (3): 582-600.(in Chinese)
[10] Rau L F. Extracting company names from text[C]// IEEE Conference on Artificial Intelligence Application. IEEE Computer Society, 1991:29-32.
[11] Ji Yuan, Xie Dong, Zhou Siming, et al. Construction method of semantic search system in power domain. [J]. Journal of Computer Systems and Applications. 2016, 25 (4):91-96.(in Chinese).
[12] Huang Y, Zhou X. Knowledge model for electric power big data based on ontology and semantic web [J]. CSEE Journal of Power & Energy Systems, 2015, 1 (1):19-27.
[13] Lao N, Mitchell T, Cohen W W. Random walk inference and learning in a large scale knowledge base [C]// Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July.
[14] WANG Zhi, XIA Shi-xiong, NIU Qiang. Research on natural language query rewriting for ontology-based knowledge base [J]. Microelectronics & Computer, 2009, 26 (8): 137-139.(in Chinese).
[15] LIU, Jingfang, ZOU Ping, ZHANG Pengzhu, et al. Research on an Improved Algorithm of Concept Semantic Similarity Based on Ontology [J]. Journal of Wuhan University of Technology, 2010, 32 (20):112-127 (in Chinese).