Research on Hybrid Index Method of Double-Layer B+ Tree for Power Big Data Considering Knowledge Graph

Research on Hybrid Index Method of Double-Layer B+ Tree for Power Big Data Considering Knowledge Graph

LingChao Gao[1], LiMing Yao[1], ZhiWei Yang[1] and Fei Zheng[2,*]

[1]State Grid Corporation Big Data Center, Beijing, 100031, China
[2]Beijing China-Power Information Technology Co., Ltd, Beijing, 100089, China

[*]Corresponding author's e-mail: zhengfei202011@163.com

Abstract. Efficient power data access can not only ensure the normal operation of power system, but also be one of the key supporting technologies to improve the operational efficiency of power grid enterprises. Aiming at the problem that traditional B+ tree does not take into account the relevance of power information fragments and cannot be indexed flexibly, a hybrid index structure and index method of power big data with knowledge graph is proposed, that is, the first layer uses B+ tree structure to store attributes, and the second layer uses knowledge graph to store attribute value, thus realizing the relevance index of power data. The simulation results show that the hybrid index structure can achieve the index of power big data with relevance information without affecting the index efficiency, which provides technical reference for efficient retrieval of multi-modal data in the future.

1. Introduction

With the continuous deployment of China's "new infrastructure" decision-making, the power infrastructure is constantly improving, the power consumption base is also growing, and a large number of electrical equipment are put into operation, which not only increases the complexity of working conditions, but also brings a large amount of data information. Due to the large number of acquisition equipment and short acquisition time interval in power system operation, the data scale is large and the data access efficiency is low. So, how to read and retrieve data quickly is a problem that needs to be solved in the current power grid data access system[1-7]. Therefore, it is the premise and foundation for the application and popularization of data services to solve the problems of long indexing time and low indexing efficiency in the process of data access.

Aiming at the problem of low efficiency of power data index the low efficiency of power data indexing, research institutions and scholars at home and abroad have made different

explorations[8-15]. At present, the index structure of B-tree or B+-tree is mainly used in power data index technology, but these index technologies have their inherent defects. B-tree is a dynamically adjusted balanced tree, each node of which can store multiple indexes, which can effectively reduce the number of I/O lookups, but is not conducive to a large number of range-based lookups. B+ tree is an improvement of B-tree. The leaf nodes of B+ tree are connected together in sequence by pointers, and the range data can be searched quickly. However, it has the problem of index node splitting, and frequent node splitting takes a lot of time. In addition, some researchers have proposed a double-layer B+ tree index structure with inverted index and B+ tree index. However, this index structure does not fully consider the correlation between power data information fragments, and can only input and output with fixed rules, but can not be indexed flexibly. Knowledge graph technology is a structured semantic knowledge base, which can fully mine the relationship between power data, establish the graph index structure within and between data sets, improve the index efficiency and realize the relevance analysis of power data at the same time, which is one of the current research hot-spots [16-19].

Therefore, based on the existing research results, this paper improves the original double-layer B+ tree structure and proposes a double-layer B+ tree index structure including knowledge graph. This structure not only meets the needs of indexing large-scale and diverse types of data, but also takes into account the relevance between power information segments, and can be flexibly indexed.

2. Power big data and its access technology

2.1. Power big data

Power big data is a relatively broad concept. At present, with the advancement of digital "new infrastructure" of power system, more and more data have been accumulated by power grid companies and their auxiliary service agencies. Thanks to various data acquisition methods, the current equipment ontology monitoring data, operating condition inspection data, environmental meteorological data, family quality history and other data are becoming increasingly large-scale. Data show that only the annual output of smart meter data of State Grid Corporation reaches 200TB, and the data volume of State Grid Disaster Recovery Center is even closer to 15PB[20]. At the same time, it meets five characteristics of big data, namely, large data volume, fast processing speed, multiple data types, great value, high accuracy and so on. Therefore, electric power big data gradually tends to form a panoramic, large and diverse comprehensive data body. How to establish a complete electric power big data processing system model integrating information collection, transmission, conversion, storage, integration, mining analysis, display, equipment maintenance and information relearning by adopting appropriate data processing principles is a major problem facing the current researchers[21].

2.2. Data access

Data access is a process of interaction between system and data. In the process of landing related big data platform in power industry, data access is an essential key link. In the face of various sources and types of data, it is necessary to access data, that is, integrate these scattered data into a unified big data platform. The data considered in power grid data access mainly include: collected measurement data, external data, primary deployment system data and so on; from the perspective of data types, data

access mainly includes structured data access, semi-structured data access and unstructured data access. From the point of view of data update time of view of data update time, access data sets are mainly historical data and real-time incremental data.

Integration and access of power grid data is the basis of subsequent data processing and data application. However, the explosive growth of massive power grid data puts forward higher requirements for the hardware environment and software system of power grid full-service data access. How to quickly read and retrieve data is an urgent problem to be solved in the current power grid data access system.

3. Data indexing technology

3.1. Indexing technology

Data indexing is a decentralized storage structure created to speed up the retrieval of data rows in a table. Index is established for tables, which is composed of index pages other than data pages, and the rows in each index page contain logical pointers to speed up the retrieval of physical data. Index technology is one of the key technologies of modern information retrieval, search application and data mining, and it is an effective mechanism to improve the efficiency of data retrieval and query.

B+ tree is a commonly used index structure in power data index, which is an improvement of B-tree. Figure 1 shows the general index structure of B+ tree:

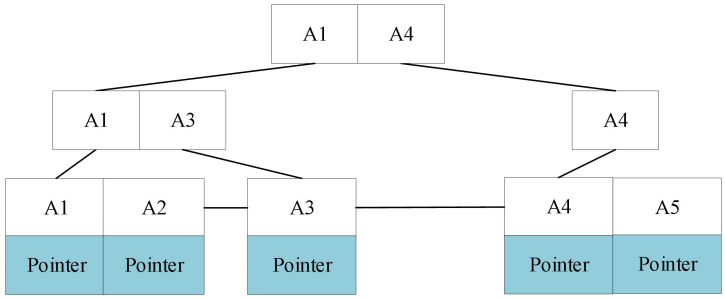


Figure 1. B+ tree index structure.

However, the index node splitting problem in B+ tree insertion greatly reduces the index efficiency, and because B+ tree can only traverse a certain path from the root node to the leaf node in the query process, it can not be well applied in information fragment association analysis. Therefore, it is necessary to introduce new data processing technology.

3.2. Knowledge graph

Knowledge graph is to organize a large amount of collected data into a knowledge base that can be processed by machines, and realize visual display. Knowledge graph is essentially a large-scale semantic network, which aims to describe various entities or concepts and their relationships in the real world. It constitutes a huge semantic network diagram, with points representing entities or concepts and edges consisting of attributes or relationships. Ternary is a basic representation of knowledge graph.

The construction of knowledge graph starts from the most original data (including structured,

semi-structured and unstructured data), and adopts a series of automatic or semi-automatic technical means to extract knowledge from the original database and the third-party database and store it in the knowledge base in the form of triples. In the established knowledge base, the points with relations are connected together by means of relations or attributes. Then, according to the storage mode of the knowledge base, the data satisfying the conditions or the description information of the specified data can be obtained through a series of query statements.

Therefore, the point-related characteristics of knowledge graph can make up for the association analysis characteristics that B+ tree index structure does not have.

4. Double-layer B+ tree indexing method considering knowledge graph

4.1. Index structure

This paper studies the B+ tree index structure and the characteristics of knowledge graph index, and designs a double-layer B+ tree index structure combined with knowledge graph according to the characteristics of power system data, as shown in Figure 2.

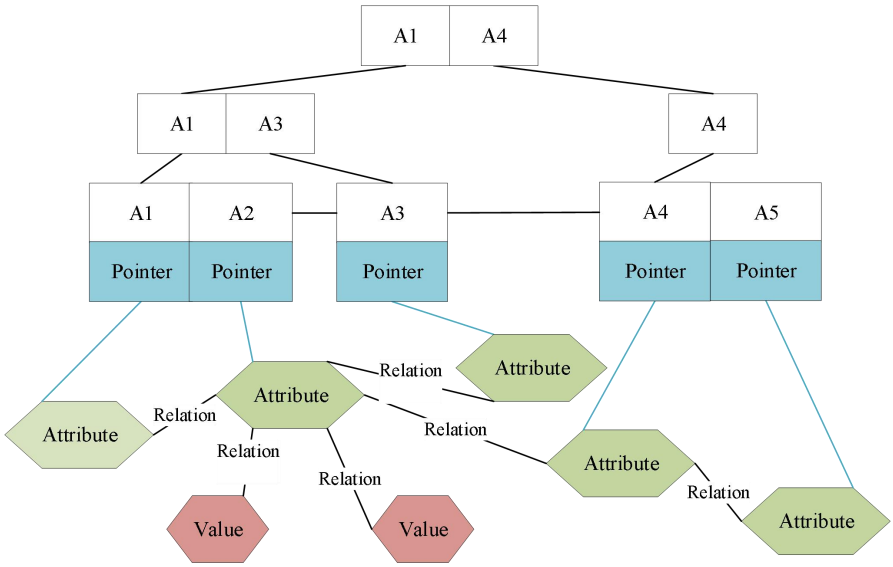


Figure 2. Schematic diagram of double-layer B+ tree index structure considering knowledge graph.

The first layer uses B+ tree to store data attributes, and the second layer uses knowledge graph to store attribute values.

B+ tree has internal nodes and leaf nodes, and the specific attributes exist in the leaf nodes of B+ tree. The leaf node contains two parts of information < Ai, Pointer>:Ai is the number of power data attributes in the index dictionary, $i \in [1,n]$, and n is the total number of attributes; Pointer is a pointer to the corresponding attribute in the second-level knowledge graph.

Knowledge graph is composed of <attribute-relation-attribute> and <attribute-relation-attribute value>. When a point represents an attribute value, it contains three parts of information. The specific structure is shown in Figure 3:

Attribute value	File number	File path
Val	Dn	Dl

Figure 3. Point structure diagram of knowledge graph.

Where *Val* represents the attribute value, that is, the content to be queried; *Dn* indicates the file number where the content is located, and each file number is unique; *Dl* is the location information of the file where the content is located.

4.2. Inquiry method

The double-layer B+ tree indexing method considering knowledge graph is shown in Figure 4.

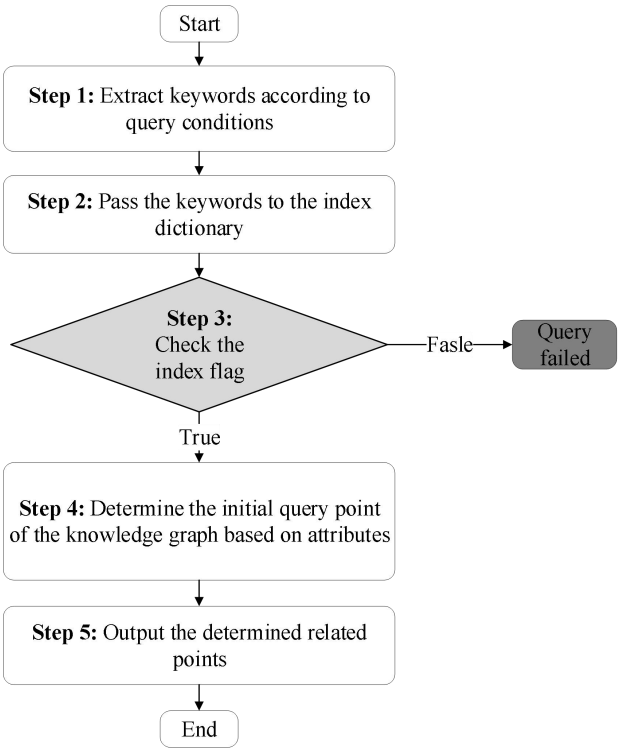


Figure 4. Flowchart of indexing process.

When power data need to be queried, keywords are first extracted according to query conditions, and then the keywords are handed over to the index dictionary. If the index flag bit is False, it means that there is no searched data in the current index and the query fails. If the index flag bit returns True, the index starting point of the lower knowledge graph is determined according to the pointer content of the leaf node after the leaf node is determined according to the upper B+ tree. Finally, query the corresponding points of keywords and a series of related points in the knowledge graph and output them.

5. Simulation experiment analysis

In this paper, the query efficiency and query results of double-layer B+ tree structure combined with knowledge graph are analyzed, and single-layer B+ tree structure is selected as a comparison.

Description of simulation example:The order of the single-layer B+ tree and the upper-layer B+

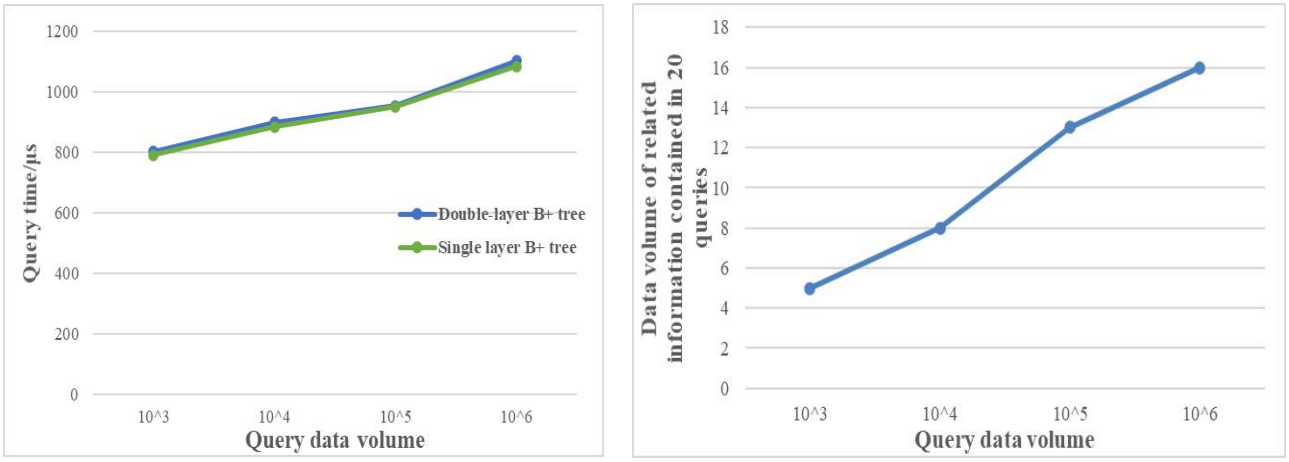
tree in the double-layer hybrid structure are set to be 8, and four experiments are carried out on the basis of these two index structures, and random content is queried 20 times each time, and the number of sample database records queried each time is 10^3,10^4,10^5,10^6 respectively. Considering the influence of data types on index efficiency, the sample database includes both character data and numerical data. Other conditions remain unchanged during the experiment.

The composition of experimental data set is shown in Table 1:

Table 1. Data set description.

Number of experiments	Total data volume	Character data amount	Numerical data quantity
1	10^3	$5*10^2$	$5*10^2$
2	10^4	$5*10^3$	$5*10^3$
3	10^5	$5*10^4$	$5*10^4$
4	10^6	$5*10^5$	$5*10^5$

By incorporating the knowledge graph method, the results of the hybrid index method of double-layer B+ tree constructed in this paper are shown in Figure 5.



(a) Comparison of query time.

(b) Hybrid B+ tree index results considering knowledge graph.

Figure 5. Hybrid B+ tree index results considering knowledge graph.

Generally speaking, compared with the original single-layer B+ tree, the double-layer B+ tree index method combined with knowledge graph can improve the query efficiency, and can find related attributes and attribute values when querying a certain attribute value. Therefore, the query results can show the power grid operation status in many aspects, provide reference for power grid workers, and ensure the safe operation of the power grid system to a certain extent.

6. Summary

There must be related power information fragments in large-scale multi-source and heterogeneous power data. Combining B+ tree and knowledge map structure can improve the efficiency of data indexing and consider the data correlation characteristics, and mine new knowledge from power data to a certain extent, and at the same time provide technical reference for efficient retrieval of multi-modal data in the future of data indexing, and at the same time, consider the data correlation

characteristics, and mine new knowledge from power data to a certain extent, at the same time, provide technical reference for efficient retrieval of multi-modal data in the future. However, how to build a scientific and effective knowledge graph in the huge power data and how to make better use of the association results found through the double-layer B+ tree need further exploration and discussion.

References

[1] LIU Bo, SHI Meng. Analysis of user access scheme of intelligent distribution network based on big data technology[J]. *Electromechanical Information*, 2017(09): 21-22.
[2] WANG Wei, LIU Yin, YU Zhanpeng, et al. System design of the big data center architecture in electric power big data environment[J]. *Power Information and Communication Technology*, 2016, 14(01): 1-6.
[3] ZHANG Dongxia, MIAO Xin, LIU Liping, et al. Research on development strategy for smart grid big data[J]. *Proceedings of the CSEE*, 2015, 35(01): 2-12.
[4] PENG Xiaosheng, DENG Diyuan, CHENG Shijie, et al. Key technologies of electric power big data and its application prospects in smart grid[J]. *Proceedings of the CSEE*, 2015, 35(03): 503-511.
[5] LIU Daowei, ZHANG Dongxia, SUN Huadong, et al. Construction of stability situation quantitative assessment and adaptive control system for large-scale power grid in the spatio-temporal big data environment[J]. *Proceedings of the CSEE*, 2015, 35(02): 268-276.
[6] SONG Yaqi, ZHOU Guoliang, ZHU Yongli. Present status and challenges of big data Processing in smart grid[J]. *Power Grid Technology*, 2013, 37(04): 927-935.
[7] Mladen Kezunovic, Pierre Pinson, Zoran Obradovic, et al. Big data analytics for future electricity grids[J]. *Electric Power Systems Research*, 2020, 189: 106788.
[8] SHI En, GU Daquan, FENG Jing, et al. Research and improvement of B+ tree indexing mechanism[J]. *Computer Application Research*, 2017, 34(06): 1766-1769.
[9] LIANG Junjie, XIAO Yao, YU Dunhui. Research and improvement of B+ tree indexing mechanism[J]. *Computer Application Research*, 2016, 33(03): 706-710+715.
[10] Amr A. Munshi, Yasser A.-R. I. Mohamed. Big data framework for analytics in smart grids[J]. *Electric Power Systems Research*, 2017, 151: 369-380.
[11] ShyamR., Bharathi Ganesh H.B., Sachin Kumar S., et al. Apache spark a big data analytics platform for smart grid[J]. *Procedia Technology*, 2015, 21: 171-178.
[12] Kirk McKusick, Sean Quinlan. GFS: evolution on fast-forward[J]. *Communications of the ACM*, 2010, 53(3): 42-49.
[13] Jeffrey Dean, Sanjay Ghemawat. MapReduce: simplified data processing on large clusters[J]. *Communications of the ACM*, 2008 ,51(1): 107-113.
[14] CAO Ju, YIN Zhe. Clouds search optimization[J]. *Computer Engineering and Science*, 2011, 33(10): 120-125.
[15] ZHANG Limei, LIU Lu. Fast search of multi-channel mass data based on optimal entropy matching[J]. *Journal of Southwest Jiaotong University*, 2012, 47(02): 313-317.
[16] QIAO Ji, WANG Xinying, MIN Rui, et al. Framework and key technologies of knowledge-graph-based fault handling system in power grid[J]. *Proceedings of the* CSEE, 2020, 40(18): 5837-5849.
[17] GAO Haixiang, MIAO Lu, LIU Jianing, et al. Review on knowledge graph and its application in power systems[J]. *Guangdong Electric Power*, 2020, 33(09): 66-76.
[18] WANG Yuan, PENG Chenhui, WANG Zhiqiang, et al. Application of knowledge graph in full-service unified data center of national grid[J]. *Computer Engineering and Application*, 2019, 55(15): 104-109.
[19] TAN Gang, CHEN Yu, PENG Yunzhu. Hybrid domain feature knowledge graph smart question answering system[J]. *Computer Engineering and Application*, 2020, 56(03): 232-239.

[20] XUE Yusheng, LAI Yening. Integration of macro energy thinking and big data thinking part one big data and power big data[J]. *Automation of Electric Power Systems*, 2016, 40(01): 1-8.
[21] MENG Xiaofeng, CI Xiang. Big data management: concepts, techniques and challenges[J]. *Journal of Computer Research and Development*, 2013, 50(01): 146-169.