Research on intelligent construction data resource catalogue based on power grid business

Lingchao Gao[a], Xiangzhou Chen[a], Shunhua Liu[b], Fei Zheng[*b]
[a]State Grid Big Data Center, Beijing 100031, China; [b]Beijing China-Power Information Technology Co., Ltd., Beijing 100089, China

ABSTRACT

With the high-quality development of informatization, power grid enterprises promote the steady development of various businesses by virtue of strong technical barriers. In this paper, it is urgent to carry out research on the construction of intelligent data resource catalogue based on machine learning technology for the sake of consolidating the data application foundation of power grid enterprises, facilitating the potential value of data and optimizing the intelligent level of data services rapidly. Additionally, in this work, we analyze the framework, construction process and key technologies of intelligent data resource catalogue. Ultimately, this work realizes the intelligent construction of data resources in power grid enterprises.

Keywords: Power grid enterprises, data application, data resource catalogue, intelligent

1. INTRODUCTION

The steady improvement of the informatization degree of power grid enterprises benefits from the iterative optimization of network technology and information technology. And the enterprise data is also growing massively with the increase of information system[1,2]. The rapid introduction of new business further aggravates the dynamic adjustment of table mode, which leads to problems such as difficult data identification, difficult data maintenance, difficult data application, difficult data front and rear correlation and so on. In order to make the construction of power grid enterprise information management platform truly take data as the core and technology as the driving force. we should rebuild the business core to create a more scene and intelligent business environment, and realize the construction of intelligent data resource directory.

In the interest of synchronize the instructions of the No. 14 document “notice on strengthening data management” of the Internet department, promote the company’s key tasks, develop the company’s data inventory, optimize the data management progress, and form a data resource directory[3,4]. Firstly, this paper studies and analyzes the development status of data resource directory construction at home and abroad, and finds that there is an optimized situation. And then, based on the response and adjustment ability of the information system, this paper develops the intelligent construction and application research of the data resource directory of power grid enterprises under the new situation and put forward the design scheme of the intelligent data resource directory, to enable the front-end business application ability and promote the effective innovation ability. Finally, the domestic and foreign demand will be improved, and the goal of improving internal quality and efficiency and sustainable external development will be achieved.

2. RELATED WORK

2.1 Machine learning technology

With the development of artificial intelligence technology, artificial intelligence related products have covered many fields such as power, finance, rail transit, smart city, education and so on[1], gradually changing people’s way of life and work. As the core technology of artificial intelligence, machine learning has become one of the most cutting-edge research directions[2,3].

Based on machine learning techniques such as decision tree and Apriori algorithm, this paper studies the construction method of data resource catalog. In 1985, Breiman put forward the CART algorithm, and proposed the use of tree

* zhengfei202011@163.com

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125064D
 · 0277-786X · doi: 10.1117/12.2662654

structure algorithm to disperse data for the first time[4]. In 2009, Li proposed a decision tree induction method based on large margin heuristics, which significantly optimized the generalization ability of decision trees[5]. In 2018, Elaidi proposed a new clustering algorithm based on two effective monitoring technologies,binary decision tree and support vector machine[6]. In 2020, Rochmawati used J48 and Hoeffding decision tree to transform clinical symptom data set into decision tree or decision rule, and classify clinical symptoms to judge whether patients are infected with Corona novel Coronavirus[7]. In terms of Apriori algorithm. In 2018, Luna proposed a maximal AprioriMR algorithm based on the previous algorithm to solve the problem of data homogeneity and regular item set extraction. This method is used to mine the compressed representation of frequent patterns, but this algorithm is only applicable to pattern mining of large data sets[8]. In 2021, Von Rueden et al. proposed an informed machine learning classification framework, which can explicitly integrate additional prior knowledge into machine learning.This method can not only improve the utilization rate of prior knowledge, but also can significantly improve the robustness of the model in the case of insufficient model training samples[9]. In 2022, Cheng et al. used Apriori algorithm to optimize XGBoost prediction model for the root alarm problem of power communication, The optimization results were significant as the data complexity was significantly improved, and the prediction accuracy of the model also achieved an ideal improvement[10].

2.2 Data management

Data management technology occupies a self-evident position in the development of information society, which is related to the development of scientific research and decision-making management. Its existence maintains the development of related systems, such as management information system, automated office system and decision support system. At present, the development of traditional data management technology lags behind and conflicts with the needs of data management and data processing, so a new data management system emerges at the historic moment[11].

In 2016, Petrasch proposed in Reference[12] that the cloud storage center (CSH) should be used to integrate different data sources and provide enterprises with an API for enterprise information management with real-time and consistent data. In 2019, Cui of Renmin University of China studied the application of data management technology in machine learning based on the application scenario of machine learning using data sets for model training[11]. In 2019, Lwin, aiming at the application of big data analysis technology in disaster management, constructed a set of spatial analysis tools of geographical visualization through data analysis and Internet of Things technology, which are used to improve spatial thinking and planning process in disaster management, so as to reduce the loss of life and property[13]. In 2021, Zainab proposed A method for centralized management of smart grid data in view of data management problems in smart grid, and provided auxiliary decision-making role for grid operation and construction through data management method[14]. In 2021, Dong Wang, a science and technology project management data model is put forward, which pertaining to the project units, themes, and between inner link. And put forward a kind of science and technology management data query method based on the relationship between the engine, which provides data query with a query method based on the relationship between the engine and improve the management of science and technology of data query efficiency[15].

2.3 Data resource catalogue

In foreign countries, there is no exact concept related to data resource directory. The existing research shows that only relevant applications based on big data technology have been carried out. For example, the “troika” disclosed by Google, the first is the distributed file system GFS running through the redundant storage mechanism, the second is the parallel processing framework MapReduce, and the third is the data storage model BigTable. The essence of the “troika” is “massive data” processing technology, which plays a key role in the integrated development of Hadoop, a distributed system infrastructure. In 2012, global data processing broke through the bottleneck and achieved a qualitative leap, and the original open-source Hadoop ecosystem officially began commercialization[16].

In China, Alibaba broke through industry barriers and took the lead in proposing the strategic development concept of a catalog of digital data resources. It created data stratification and horizontal decoupling in R&D, and accumulated public data capabilities[17] and data analysis capabilities. At the end of 2015, Didi launched the data resource catalog strategy to build a travel big data resource catalog. Through the self-generated platform, it realized service-oriented, configuration-oriented, dynamic configuration of the business side, and realized the support for the rapid development of new business[18], which solved the original problem. There are problems such as poor platform sharing and repeated construction.

Although domestic advanced Internet enterprises have made phased achievements in data resource catalogue, which can provide theoretical and practical reference for the application of data resource catalogue in power business. But there is a big difference between Internet business model and power business. In order to realize the intelligent construction of data

resource of power grid enterprises, the architecture design and application research of data resource directory should be carried out according to the actual business and demand of power industry.

2.4 Existing work basis of power grid enterprises

The preliminary work provides a solid foundation for the development of the RESEARCH and development, concentrated in data integration, analysis, application and other modules, mainly reflected in data center construction, data resource management tool construction, master data construction, data sharing and business integration construction and the construction of unified data center analysis domain construction.

In 2016, the state Grid all-service Unified Data Center proposed the construction goal of "managing enterprises with data and driving businesses with information". After three years, the expected experimental results have been basically realized. This work has created a standardized and unified model, clean and transparent data, flexible and intelligent analysis, and contributed to the birth of an operating environment with complete data resources and efficient computing and analysis capabilities. It has become an integrated hub of the company's all-business, all-type and all-time data. In 2019, in order to strengthen the data management of State Grid Corporation of China, accelerate the establishment of data management system, promote the Shared data fusion applications, give full play to the value of data, State Grid Corporation of China organized the revision of *Data Management Methods of State Grid Corporation of China*.This work points out the method to establish a unified management, grading is responsible for data management mechanism to promote the data management system to the ground. In 2020, the Internet Department of the national power grid co., LTD studied and formulated *the Key Work Arrangement for Data Management* to strengthen the special improvement of data standards and quality, promote key breakthroughs in data opening and service application, give full play to the important role of data as a factor of production, and implement the strategy of service companies.

To sum up, data management has become one of the important strategies for the development of State Grid Corporation of China. And the construction of data resource catalog can realize classified storage of all kinds of data in the database, which is the key to realize efficient data management. In order to promote the construction of state Grid data resource catalogue, this papercarries out the research on the construction technology of intelligent data resource catalogue based on power grid business by using learning technologies such as decision tree and Apriori algorithm.

3. INTELLIGENT DATA RESOURCE CATALOGUE DESIGN

3.1 Construction framework of intelligent data resource directory

Firstly, Classify the types of data objects managed in the resource catalog and sort out the attributes of information items needed by each type of data objects. And according to each type of information attribute to implement collection, automatic acquisition, intelligent analysis, classification and association analysis. The overall research framework is shown in Figure 1.

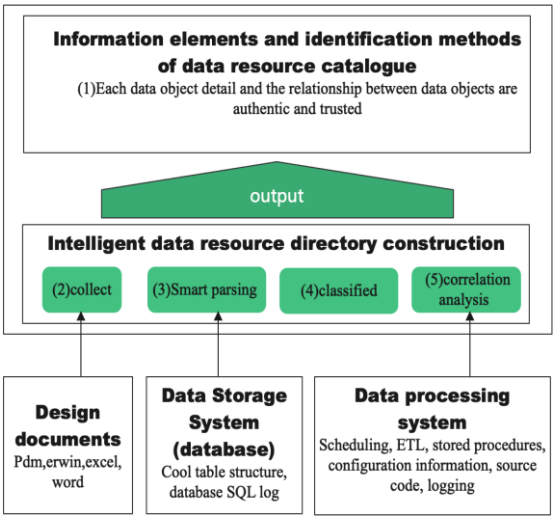

Figure 1. Framework for building an intelligent data resource directory.

By collecting, intelligently analyzing, classifying and associating the information of design document, database and data processing system, a unified data resource catalogue is constructed, and the information elements and identification methods of data resource catalogue are output.

3.2 Construction steps of intelligent data resource catalogue

Combined with the objectives and requirements of inventory of data resources of State Grid Corporation of China, it is necessary to sort and analyze the company's data resources to show the significance of the data resource catalogue and provide technical metadata and business metadata. Combined with the practice of the government and enterprises in the industry, the method of constructing the data resource catalogue in the state grid system is designed, and the overall steps of the study are shown in Figure 2.

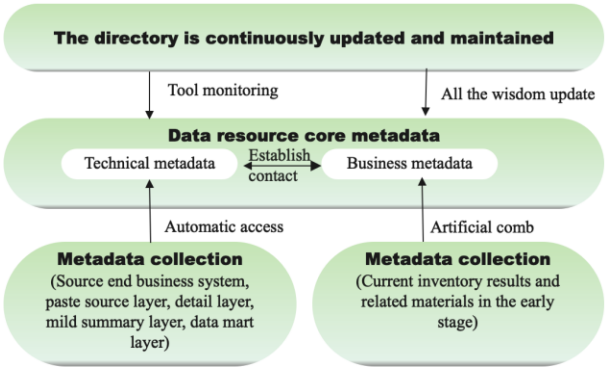

Figure 2. Steps for building an intelligent data resource directory.

The method of constructing data resource directory is divided into three steps. The first step is to carry out metadata collection by means of automatic extraction and manual sorting. Metadata collection is an important hub for establishing the association between Technical Metadata and business metadata, and its main function is to provide data for the construction of data resource directory. The second step is the establishment of metadata association relationship, that is, the association relationship between Technical Metadata and business metadata is established by using artificial intelligence method. The third step is to use Zhongzhi's metadata change detection and update tools to continuously update the data of resources.

3.3 Construction process of intelligent data resource catalogue

The construction process of intelligent data resource catalog construction is shown in Figure 3, including four steps: the first step is to obtain technical metadata; The second step is to pass business metadata; The third step is to improve the correlation between technical metadata and business metadata. The fourth step is to update the data resource catalog.

3.3.1 Technical Metadata Extraction. (1) Online configuration of database information of system layer, paste source layer, sharing layer and application layer through tools. (2) Tools are used to automatically extract technical metadata of all layers, including the data entities (library, table, field) in the process and the logic in the processing of data entities, and realize the automatic update of normalization.

3.3.2 Combing Service Metadata. As there is no specific calculation definition standard for service metadata, it is needed to manually sort the service metadata and import the sorted service metadata.

3.3.3 Associating Technical Metadata with Service Metadata. The data resource catalog support tool is used to associate technical metadata information with business metadata information. It reflects the business characteristics of the data directory to facilitate users to understand and use data[6].

On the basis of effective metadata management, the data resource catalog adds related data processing process, personnel, platform and quality related capabilities, including intelligent analysis of data, data classification processing and table association analysis.

3.4 Research technology of intelligent data resource catalogue

The construction of intelligent data resource catalog is mainly based on machine learning to realize the automatic

construction of data resource catalog, and the construction process is consistent with that of the association technology of technical metadata and business metadata.

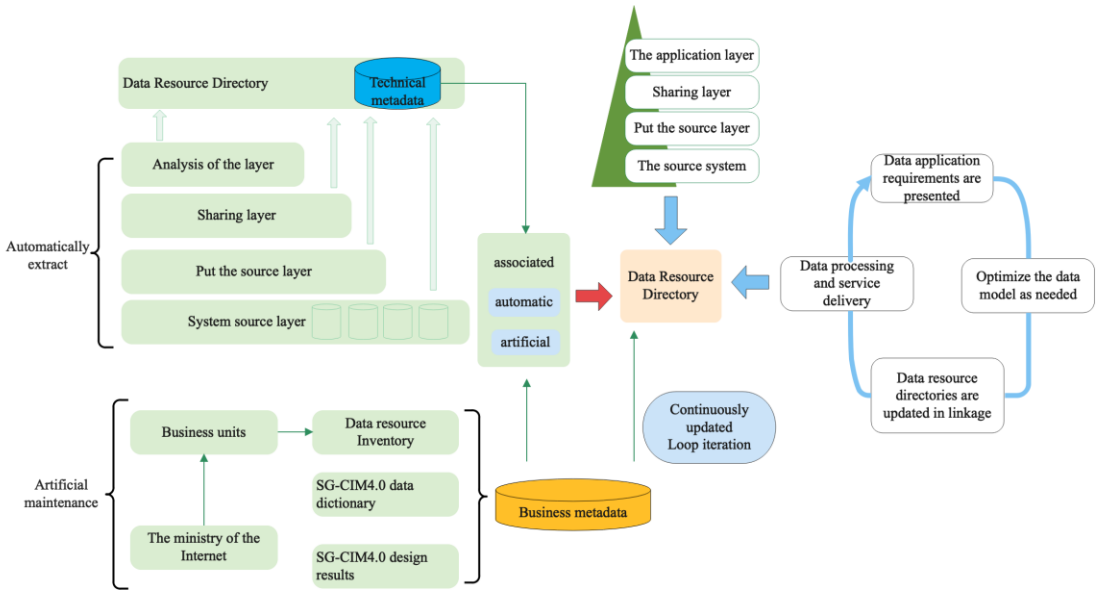

Figure 3. Intelligent construction process of the data resource catalog.

3.4.1 Intelligent Data Analysis Technology. According to the data analysis standard, the built-in data standard is set up to realize the automatic recognition of data format and technology type. Through natural language processing, feature analysis and other methods, the semantic content of scanned metadata is identified and the business types of fields are analyzed,that help users quickly and deeply understand the meaning of data representation. The function of sorting and filtering are supported by field name, table name and business type, and identification results can be modified according to the matching rate of identification results and sample data of fields[17]. The intelligent analysis model in this paper innovatively adopts supervised learning and naive Bayes algorithm in metadata semantic recognition. The results of intelligent data analysis can help users generate a management tool like a data dictionary, and provide a foundation for subsequent data association analysis and table business analysis.

3.4.2 Data Classification Technology. The construction of decision tree is binary tree, which has different meanings in different situations. In the case of category identification represented by leaf nodes, it is called classification tree; In the case where constants are assigned to each leaf node, also known as regression trees, piecewise constants represent regression functions. Note that when it is a classification tree, multiple leaves may have the same identity.

According to the characteristic that decision tree can deal with irrelevant characteristic data, this paper selects decision tree algorithm to classify different types of data. The global training data is placed at the root node for splitting. At each node, optimization criteria (such as optimal splitting) are determined based on some basic principles. According to the initial and alternative split points, all data are divided into left and right children, and then continue to split left and right children, and finally stop computing at a node, so as to achieve data classification. If the algorithm stops at a node, the following may occur.

(1) The tree depth reaches the specified maximum value;

(2) The number of training samples is constrained and less than the specified value;

(3) when the global samples are of the same class;

(4) Compared with random selection, the best split that can be selected has reached the fitting state.

3.4.3 Table Association Analysis Technique. Table association analysis technology optimized the complexity of Apriori algorithm, mainly reflected in the number of scanning database, the number of waiting option set andstorage space improvement. A compression matrix is introduced into The Apriori algorithm. According to the properties of frequent

item sets, the Apriori algorithm seeks frequent item sets through layer-by-layer iterative search: that is, further searches k item sets through (K-1) item sets.

Apriori algorithm based on compression matrix greatly reduces the computing power. In the process of complete compression of the matrix, when K increases, the matrix becomes smaller, so the workload of calculating the support degree of K term sets is correspondingly reduced and the calculation efficiency is improved. The optimized algorithm first processes the data and finally synchronizes it to a Boolean matrix data table. In the source database, "0" and "1" represent data. Compared with the original data, the storage space occupied by the data is much less than that of the original data, thus reducing the space occupancy.

4. CONCLUSION

In this paper, we present the machine learning technology domestic and overseas, which researches the status of data resource catalog construction. Our method integrates the current data application situation of power grid enterprises. We leveraged the construction of intelligent data resource catalog of power grid and analyzed the construction framework, construction process and related technologies of intelligent data resource catalog. In addition, we put forward data intelligent analysis, data classification and table association technology innovatively. To this end, our approach optimizes the intelligence of the data resource catalog and resolves the analysis and construction work.

REFERENCES

[1] Ross, M., Graves, C. A., Campbell J. W. and Kim, J. H., "Using support vector machines to classify student attentiveness for the development of personalized learning systems," 2013 12th International Conference on Machine Learning and Applications, 325-328 (2013).

[2] Li, F. Z., Zhang, L., Yang, J. W., Qian, X. P., Wang, B. J. and He, S. P., [Lie Group Machine Learning], China University of Science and Technology Press, Hefei, China, (2013). (in Chinese)

[3] Li, F. Z., Zhang, L. and Zhang, Z., [Lie Group Machine Learning], https://doi.org/10.1515/9783110499506, (2019).

[4] Breiman, L., Friedman, J., Olshen, R. A., et al., [Classification and Response to Criticismsion in Trees], Wadsworth, Belmont, (1984).

[5] Li, N., Zhao, L., Chen, A. X., Meng, Q. W. and Zhang, G.-F., "A new heuristic of the decision tree induction," 2009 International Conference on Machine Learning and Cybernetics, 1659-1664 (2009).

[6] Elaidi, H., Elhaddar, Y., Benabbou, Z. and Abbar, H., "An idea of a clustering algorithm using support vector machines based on binary decision tree," 2018 International Conference on Intelligent Systems and Computer Vision (ISCV), 1-5(2018).

[7] Rochmawati, N., et al, "Covid symptom severity using decision tree," 2020 Third International Conference on Vocational Education and Electrical Engineering (ICVEE), 1-5 (2020).

[8] Luna, J. M., Padillo, F., Pechenizkiy, M. and Ventura, S., "Apriori versions based on mapreduce for mining frequent patterns on big data," in IEEE Transactions on Cybernetics 48(10), 2851-2865 (2018).

[9] von Rueden, L., et al., "Informed machine learning—A taxonomy and survey of integrating prior knowledge into learning systems," The IEEE in the Transactions on Knowledge and Data Engineering, doi:10.1109/TKDE, 3079836 (2021).

[10] Lu, M. C., Lou, P., Zhu, J., Li, L., Cui, X. and Sun, Y., "Power communication network root alarm prediction based on apriori-bayesian optimization XGBoost," Electric Power Construction 43(01),113-121 (2022).

[11] Cui, B., Gao, J., Tong, Y., Xu, J., Zhang, D., and Zou L., "Progress and trend of new data management systems," Journal of Software 30(01), 164-193 (2019).

[12] Petrasch, R. and Hentschke, R., "Cloud storage hub: Data management for IoT and industry 4.0 applications: Towards a consistent enterprise information management system," 2016 Management and Innovation Technology International Conference (MITicon), MIT-108-MIT-111 (2016).

[13] Lwin, K. K., Sekimoto, Y., Takeuchi, W. and Zettsu, K., "City geospatial dashboard: IoT and big data analytics for geospatial solutions provider in disaster management," 2019 International Conference on Information and Communication Technologies for Disaster Management (ICT-DM), 1-4 (2019).

[14] Zainab, A., Ghrayeb, A., Syed, D., Abu-Rub, H., Refaat, S. S. and Bouhali, O., "Big data management in smart grids: Technologies and challenges," IEEE Access 9, 73046-73059 (2021).

[15] Wang, D., Zhang, Z., Xu, C. and Wang, Z., "Data query method of science and technology management based on relational engine," 2021 IEEE 6th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), 111-115 (2021).

[16] Zhu, H., "Create enterprise data resource directory to promote enterprise intelligent operation," Communications Enterprise Management, (2018).

[17] Zhang, Z., Sun, Y., Chen, C., et al., “Power grid data quality verification method and verification system based on Hadoop,” Computer Research and Development 2014(S2), 134-144 (2014).
[18] Xu, G., “Research on omni-channel scenario-based operation based on enterprise data resource catalog,” Information Communication (8), (2017).