Framework of Intelligent Analysis and Mining for Power Big Data

Chong Wang¹, a, Mingming Zhang² b, Gaopan Huang³ c, Haoxiang Dou⁴ d, Menghan Xu⁵ e

(Big data laboratory of State Grid Jiangsu Electric Power Company Limited Information and Telecommunication Branch Nanjing, Jiangsu 210024)

E-mail: ¹wangchong219@163.com ²zhangmms@163.com ³huanggpan@163.com ⁴douhx1004@163.com ⁵757660217@qq.com

Abstract. With the continuous development of smart grid, how to make full use of and analyze these data and get valuable information quickly is the key points which need to be solve in the big data processing. This paper constructs the framework of intelligent and efficient mining for power big data, analysis technology, high performance data analysis technology and the parallel analysis framework for power big data. Several service system described in detail the key techniques of intelligent analysis of large data mining and electricity from the power of big data characteristics.

1. Introduction
With the deepening construction of smart grid, at present, State Grid Corporation has built four types of data center platforms: structured, unstructured, massive historical/real-time, and grid space. The data center platform has accumulated a wealth of big data resources, effectively supporting the integration of enterprise data sharing, analysis and decision-making system. Zhang et al. pointed out that the smart grid is one of the important areas for the application of big data, and gives the key technology system framework and roadmap for big data for the smart grid [1]. At present, the State Grid Corporation has accumulated a large number of data resources for its decision-making analysis provides a good condition, but there are scattered data storage, cross-type mass data analysis, real-time decision-making is not in place and other issues.

Big data, as a means of analysis and decision-making, has drawn great attention from academia and industry. Rao et al. combined with the needs of the development of energy Internet, and explored the power big data based on the system architecture and the corresponding standard system [2]. Cheng et al. proposed a scalable collaborative big data mining platform architecture based on Apache Spark and Jupyter, which leads to higher productivity of the big data mining applications in smart grids [3]. Liu et al. proposed an architecture model of data integration in power field, a concept to build data extraction wrapper through bottom-up pattern and an automating repetitive record detection model for unsupervised learning, which can help users to obtain required data by conducting data integration for information islands, storage of unstructured data in business systems[4]. K. Vimalkumar et al. proposed a big data framework based on various machine learning techniques and detected intrusion based on the classification method on the synchronized phasor dataset. This use of feature selection and reduction algorithms reduces the amount of time it takes to predict [5]. Cheng et al. uses distributed K-means, SVM algorithms and Hadoop-based big data platform to realize low-voltage
genetic diagnosis model, which improves accuracy and reliability [6]. Chen et al. proposed the overall framework of big data platform based on safety supervision, and demonstrated the process of safely monitoring and controlling big data application structure from source to end user, which bringing new development opportunities for the construction of intelligent safety monitoring platform [7]. Maria Stefan proposed a method of building an efficient visualization system to extract the information of the smart meter, which can be used in a meaningful way. This method improves the visualization efficiency [8]. In order to better apply big data to various industries, Fan et al. put forward an improved KNN classification algorithm for large-scale data classification [9]. Li et al. proposed a big data preprocessing algorithm based on partial order attribute reduction, which greatly simplified the dimension of big data and improved the efficiency and accuracy of subsequent algorithms [10]. Wang et al. used MapReduce and Spark computing mode to solve a parallel distributed load forecasting algorithm [11]. Zhao J et al. [12] stated the concept of big data and analyzed the opportunities and challenges. Yang Z et al. [13] provided an electric power data deployment solution for distributed memory computing to solve load balance and improving performance. Huang J et al. [14] introduced a universal framework of electric power big data platform, based on the analysis of the relationships among the big data, cloud computing and smart grid. Bhuiyan S M A et al. [15] focused on data analytics to effectively interpret phasor measurement unit (PMU) data from smart grid.

In this paper, power big data as the research object, focusing on building an intelligent large-scale data analysis and mining technology system framework, and analyzes the key point and difficulties from four aspects: the characteristic analysis for big data-intensive power big data, the high performance data analysis based on memory calculation, the power data parallelization analysis framework and service architecture, and the bus ultra-short-term load forecasting technology based on data mining.

The remaining part of this paper is organized as follows: Part 2 builds an intelligent and efficient analysis and mining technology framework for power big data; Part 3 focuses on key technologies for intelligent analysis and mining of big data; Part 4 summarizes and prospects.

2. Frame of intelligent and efficient analysis and mining technology for power big data

Big data accumulated along with the continuous development of system and data service center of the complex data of smart grid formation, the analysis of the data can not only bring rich conditions, but also to many aspects of the ability to influence the computation speed, system analysis and construction mode etc. For multi-sources and heterogeneous data from the fourth major data centers of SGCC and external data sources. Framework on power big data intelligent and high efficiency analysis and mining includes the data source layer, the characteristic analysis layer, the service layer and the verification layer. The overall technical framework is shown in Figure 1.
Figure 1. Frame on intelligent efficient analysis and mining technology for power big data

The data source layer mainly describes the data source of power big data, which mainly includes two parts: (1) Structured, semi-structured and unstructured data and geographic information space data of the four major data centers in SGCC; (2) Weather, social, mobile and other external data.

The feature analysis layer mainly describes the data feature extraction method, the data model construction method and the data parallelization analysis framework.

Service layer mainly through service structure and service interface to build power big data parallel analysis service system.

Application layer using ultra-short-term bus load forecasting as the application scenario, by calling the appropriate interface to complete real-time load big data analysis.

3. Key technology analysis

This paper focuses on four key aspects of technical analysis, such as the characteristic analysis technology for computing intensive power big data, the high performance data analysis technology based on memory calculation, the big data parallel analysis framework and service system and the bus ultra-short-term load forecasting technology based on data mining.

3.1. Feature Analysis Technology for Power Big Data

In order to achieve real-time and efficient power big data analysis, the following aspects should be considered.

(1) Data model based on data objects entity and attributes

To abstract the characteristics of power big data, we first need to model all kinds of data. Along with the emergence of massive unstructured data, big data processing analysis needs to be able to provide unified services for structured, semi-structured and unstructured data. A good data model is the lowest level of capability for unified data processing and analysis of big data. Power big data contains structured, semi-structured, unstructured data, with obvious characteristics of business objects. By modeling business objects, data object entities can be generated.

Through the abstract analysis of the records of business objects and the relationships among records, it can generate attribute features such as time dimension, spatial dimension and association rules of data content, which are also the core features of power big data. Therefore, we should design a big power data model based on the above data attributes. By analyzing the other behavioral characteristics of the data object entity, the related attribute features of the object entity behavior feature are increased, taking into account the external factors such as the data generation background.
and the field to which they belong. Finally, all the attributes are classified to build a general power big data model which proposes object-oriented entities, highlights attribute features and supports complex retrieval.

(2) Big data feature extraction based on data model

The target of feature extraction of large data is to extract a subset from the correlation feature set to establish an effective prediction model. By removing irrelevant and redundant features, feature extraction can alleviate dimensional disaster, improve generalization performance, accelerate learning process and improve model interaction, thereby improving the performance of the model. At present, most of the hypothetical training examples are known for feature extraction. And the process of the whole feature extraction is carried out off-line, but this assumption is not necessarily true in reality.

For multi-domain and multi-source heterogeneous power big data, it is necessary to study data feature extraction methods, on the basis of large power data model, using the key technologies of singularity detection, parameter training and space measurement analysis, after extracting and correcting multi-source heterogeneous data and cleaning index, linear and nonlinear methods are used to extract feature data. Finally, the spatial-temporal dimensions of power big data and the characteristics of association rules are obtained. The entire technical route shown in Figure 2.

![Figure 2. Frame of feature extraction for power big data](image)

(3) Data clustering mining and correlation mining analysis based on data features

The power big data has the characteristics of low analysis efficiency, unstructured and so on. Therefore, special processing of data must be carried out. Clustering analysis of big data from the unstructured data through a specific algorithm to attribute that sample characteristics, however, attribute vector to build a big data sample is a time-consuming process, so it requires the use of parallel computing to improve computing efficiency.

Search clusters in different subspaces using clustering algorithm of data mining clustering based on data characteristic, which has strong characteristics of dependence were divided into the same cluster, then from each cluster in choosing a subset of representative line together constitute the feature subset, in order to remove irrelevant features and redundant the purpose of feature.
In order to solve the problem of efficient clustering and correlation analysis of power big data, the clustering method based on subspace partitioning is used to extract the knowledge from high-dimensional complex data sets to support decision support. In clustering mining technology, using collaborative clustering and time evolution clustering, we fully consider the timing and spatial characteristics of big data of power; In relational mining technology, we use distributed parallel association rule mining optimization and association analysis based on data object relation to improve the efficiency and accuracy of correlation analysis. The entire technical route shown in Figure 3.

![Figure 3. Data clustering mining and correlation mining analysis based on data features](image)

(4) Big data real-time analysis and distributed storage
At present, the rapid development of large data has seriously impacted the traditional data stream storage and data flow processing. This is a real-time data record sequence. Its data recording is not only orderly but also huge amount of data. More and more applications will have to consider how to process the massive data in real-time. Real-time big data analysis refers to the analysis of huge data, making use of big data technology to finish the analysis efficiently, achieving the approximate real-time effect, and timely reflect the value and significance of data. Technologies for big data include large-scale parallel processing (MPP) database, data mining, distributed file system, distributed database, cloud computing platform, Internet and extensible storage system.

![Figure 4. Real-time analysis platform for big data](image)

The big data real-time analysis platform is mainly composed of distributed storage computing framework, real-time flow computing framework, data mining module, statistical analysis module and so on. The platform proposed in this paper has the characteristics of supporting mass data storage and computing, high reliability, high fault tolerance and so on. The performance advantage and computing advantage of the platform in mass data processing will bring greater convenience and better application experience for enterprises and users.

3.2. High performance data analysis technology based on memory computation
The list of authors should be indented 25 mm to match the abstract. The style for the names is initials. In the era of big data, the amount of data increases dramatically, and the speed of information growth is beyond imagination, which is the key factor affecting the performance of the data center. How to maintain better performance in mass data processing is a very challenging problem. Compared with traditional methods, memory computing has the advantage of giving full play to multi-core capability, processing data parallelism, and increasing the speed of memory reading. Data is stored in memory.
according to optimized column storage mode. Memory computing can make real-time analysis and operation of massive mass data without prior data preprocessing and data modeling.

In order to satisfy the demand of data analysis performance brought by intelligent and efficient power big data mining applications, this paper focuses on analyzing the memory-oriented data analysis and optimization techniques based on the existing distributed analysis framework based on Spark.

First optimization design method for transmitting data, and automated merging results by dividing the intermediate results of the data analysis, saving I / O performance overhead of the system, to further enhance the speed of analysis.

At the same time, the optimal execution plan is automatically selected based on CBO technology according to the size of the source table, the size of the middle table, the distribution of data, the data skew status and other statistical information in different business scenarios. Before performing the scheduled task, the data sources are filtered according to the existing query conditions in the business scenario. Filtering data that does not participate in actual operations in the actual process and then performing unified calculation greatly improve the operation efficiency of big data analysis.

3.3. Key points and difficulties
(1) Real-time analysis and optimization

Power big data includes the cross-specialty high-dimensional time series data such as sending, losing, changing, matching, using, adjusting and so on. It usually has the characteristics of large volume, high dimension, strong aging. It puts forward higher requirements in cross-area heterogeneous data fusion, data computing and processing capabilities, and analysis of operational efficiency. It brings new challenges to real-time data mining such as rapid cluster analysis and complex association analysis. Therefore, for real-time analysis of power big data mining technology optimization, how to achieve timing index enhanced data integration, improve memory-based data processing capabilities to enhance the efficiency of parallel algorithm analysis, correlation analysis to improve cluster performance and achieve efficient low similarity measure, it is one of the key points and difficult to be solved.

(2) Data analysis

In real-time application, data analysis techniques verify the process, the power of large data collection is often distributed over a wide range of high sampling frequency, and often suffer from data loss or distortion, burr data, thus affecting the accuracy of analysis and prediction results. In non-real-time analysis scenarios, this situation can often be identified quickly and solved by interpolation, smoothing, and more. However, real-time analysis of the scene, due to the lack of context data (mainly referring to the follow-up data) cannot be easily positioned to identify, it is difficult using traditional means to correct. Therefore, in view of the applicability of intelligent big data analysis and mining technology in real-time application scenarios, how to overcome the data quality risk based on agile data quality management, make up for the lack of contextual data through data collaborative analysis, ensure the accuracy of data analysis and prediction through the development of efficient fault tolerance mechanism, It is the key points and difficulties that need to be paid more attention and solved.

4. Conclusion

This paper focuses on intelligent power grid analysis and mining of key technical framework. Traditional power data analysis has been upgraded from mathematical statistics, hypothesis testing, statistical description and inference to intelligent algorithms, data processing, parallel computing, visualization and decision support. This paper advances the parallel design and application of data mining analysis algorithm, and provides theoretical basis and basic technical support for the efficient value mining and online decision analysis of power business data.

References
[1] ZHANG Dongxia, MIAO Xin, LIU Liping, ZHANG Yan, LIU Keyan. Research on Development Strategy for Smart Grid Big Data [J]. Proceedings of the CSEE, 2015,35(1): 2-12.

[2] RAO Wei, JIANG Jing, ZHOU Ai-hua, XIE Ruo-cheng, QIU Hong-bin. Research on Power Big Data Basic Architecture and Standard System for Global Energy Interconnection [J]. Electric Power ICT, 2016,14(4):1-8.

[3] Cheng Tianshi, Jiang Wei ,et al. Scalable and cooperative big data mining platform design for smart grid[J]. 2016 China International Conference on Electricity Distribution (CICED).2016:1-5.

[4] Donglan Liu; Lei Ma; Xin Liu; Hao Yu; Hu Tan; Xiaohong Zhao; Yong Zhao; Guodong Lv,et al. Research on key issues of data integration technology in electric power system in big data environment[J]. 2017 IEEE 9th International Conference on Communication Software and Networks (ICCSN).2017:1368-1372.

[5] K. Vimalkumar , N. Radhika , et al. A big data framework for intrusion detection in smart grids using apache spark[J]. 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI).2017:198-204.

[6] Cheng Xueting; Liu Huangming; Zheng Huiping; Song Shuyong; Wang Tan; Shi Yuxin; Shen Zeyuan, et al. Diagnosis model of low voltage generation based on large data[J]. 2017 IEEE Transportation Electrification Conference and Expo, Asia-Pacific (ITEC Asia-Pacific).2017:1-6.

[7] Yan Chen; Xin Li, et al. Research on big data application in intelligent safety supervision[J]. 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA).2017:473-477.

[8] Maria Stefan; Jose G. Lopez; Morten H. Andreasen; Rasmus L. Olsen, et al. Visualization Techniques for Electrical Grid Smart Metering Data: A Survey[C]. 2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService).2017:165-171.

[9] Fan Cunjia, Wang Yousheng, Bian Hang. Improved KNN text classification algorithm [J]. Foreign Electronic Measurement Technology, 2015,34(12):39-43.

[10] LI Gang, JIAO Pu, WEN Fushuan, SONG Yu, SHANG Jincheng, HE Yang. A Partial Order Reduction Based Method for Big Data Preprocessing in Smart Grid Environment [J]. Automation of Electric Power Systems, 2016,40(7):98-106.

[11] WANG Dewen, SUN Zhiwei. Big Data Analysis and Parallel Load Forecasting of Electric Power User Side [J]. Proceedings of the CSEE, 2015,35(3):527-537.

[12] Zhao J, Wang Y, Xia Y. Analysis of Information Security of Electric Power Big Data and Its Countermeasures[C]// International Conference on Computational Intelligence and Security. IEEE, 2017:243-248.

[13] Yang Z, Zhang C, Hu M, et al. An Electric Power Big Data Deployment Solution for Distributed Memory Computing[C]// Seventh International Symposium on Parallel Architectures, Algorithms and Programming. IEEE, 2016:155-161.

[14] Zhan J, Huang J, Niu L, et al. Study of the key technologies of electric power big data and its application prospects in smart grid[C]// Power and Energy Engineering Conference. IEEE, 2015:1-4.

[15] Bhuiyan S M A, Khan J F, Murphy G V. Big data analysis of the electric power PMU data from smart grid[C]// Southeastcon. IEEE, 2017:1-5.