

Research of Quality Management Method Based on Power Big Data

Nige Li, Min Xu, Jie Fan, Wantian Cao

State Grid Smart Grid Research Institute, Nanjing, China

* linige@sgri.sgcc.com.cn

Keywords: power big data; quality management

Abstract. Power big data is big data concepts, technologies and methods practiced in the power industry. Power big data related to electricity generation, transmission, substation, distribution, electricity, dispatching the links, cross-unit, multi-disciplinary, cross-business data analysis and mining. Big data has great potential value in the smart grid, and in order to extract its potential value, the data must be guaranteed a certain quality standards. It can provide high value-added content services to the industry by guaranteeing or improving the data quality of power big data. This paper researches the quality management method based on power big data, including an analysis of common data quality evaluation, classifies the quality of power big data, and on this basis, given the solutions of improving data quality. It shows that to research of quality management method based on power big data is meaningful and necessary.

Introduction

Power big data integrate data of production, operation, marketing, operations and management on electricity enterprise[1]. For power big data analysis and processing has become the power enterprises to deepen the application level, improve the application level, and strengthen management of powerful techniques. Power big data is not just a technological progress, but also the major changes in the entire power system in the era of big data development ideas, management system and technical route.

Power big data features can be summarized as 4 V 3 E ". 4 "V" means Volume、 Variety、 Value、 Velocity, and 3 "E" means Energy、 Exchange、 Empathy[2].

Big data not only means a large amount of data, but also represents the potential value of the large. In order to extract its potential value, data must guarantee to achieve a certain quality standards, that is, the quality of the data is the premise and basis of big data to play. According to a study of the University of Texas, a study of data validity shows that by improving the use of their data and data quality, companies can significantly improve the performance of the business. Such as the power industry, when the data utilization rate of 10%, the power industry per capita output will be able to upgrade around 18%. In addition, by upgrading the power grid data quality, ROE of power industry will be able to upgrade 218%[3].

The power industry is enough to carry out data quality management in the era of big data era from the increase of ROE of power industry. But the power industry in the era of big data to data quality management research is not only in this. The power industry can apply results of big data quality management of power network assets system to other business systems, in order to make high quality power big data and customers closely coupled to achieve precise positioning of the user, or the high quality of the electric power data and national economy tight coupling to achieve accurately restore the trend of regional economic. There are also high quality power data and power production closely coupled, to achieve feedback guidance for power facilities design and production stage. In short, it can provide high value-added content services to the industry by guaranteeing or improving the data quality of power big data.

The characteristics of power big data based on Smart Grid

The power big data of smart grid has the following features[4]:

Huge data volume. Data volume from the TB level, jumped to the PB level.

Data types are various, including real-time data, historical data, and text data, multimedia data, time series data of structured and semi structured data and unstructured data. The frequency and performance requirements of various data query and processing are also different.

There is a low value density. For example, in the condition monitoring of power transmission and transformation equipment, most of the collected data are normal data. But the rest of the abnormal data is the most important basis for the status of maintenance.

High velocity processing. Analyze the large amount of data in a fraction of a second to support decision making. The performance of online state data is much higher than that of offline data. This kind of online data analysis and mining is different from traditional data mining technology.

Common power big data quality evaluation index

Generally speaking, common power big data quality evaluation index mainly has the following five:

Correctness: The correctness of power big data is the real and accurate reflection of the data and the extent of the real world entity. The accuracy of the selected topic is selected by the consistency of the entity data value and the reference source of the correct information.

Integrity: The integrity of power big data mainly includes the entity integrity, the integrity of the integrity, the referential integrity and the integrity of the integrity.

Consistency: The consistency of power big data is usually whether the logical relationship between the relational data is correct and complete. It includes the concept of consistency, consistency and consistency range format.

Completeness: The completeness of power big data means that all the required data is there. From the vertical view, it refers to the data without duplication. From the horizontal, it refers to the key attributes of the data is not missing.

Timeliness: The timeliness of power big data refers to the big data only in a certain period of time for decision-making with the value of the attribute. Over a certain period of time, the value of the big data is reduced or lost value.

Power big data quality problem classification

Because of big data source extraction method and correctness, as well as the problem of big data center itself, there are different problems in the quality of the data center. And in different stages of data processing, it will affect the quality of data in different degree[5]. As shown in Figure 1.

Firstly, in the stage of big data acquisition, when using the manual input, due to the low accuracy and low efficiency will cause big data itself is not correct and timely. And when using automated data acquisition, but also due to the hardware and software problems will lead to problems in the integrity and credibility of big data. Because of the quality of the source data itself, it will affect the accuracy and credibility of other applications based on big data.

Secondly, in the stage of big data integration, big data quality problems in this stage is mainly due to the different data sources in big data format, structure and semantic inconsistency, and makes in data management and integration time will cause the error of the mapping relation and the data record is not complete and other issues.

Finally, in the stage of big data application, the main problem is not the correct relationship between big data and application. This problem will have a great impact on the accuracy of the application.

In addition, according to the different relationship between the data source for big data center, data quality problems can be divided into single data source and multi data source quality problems, and these problems will eventually in the example is embodied.

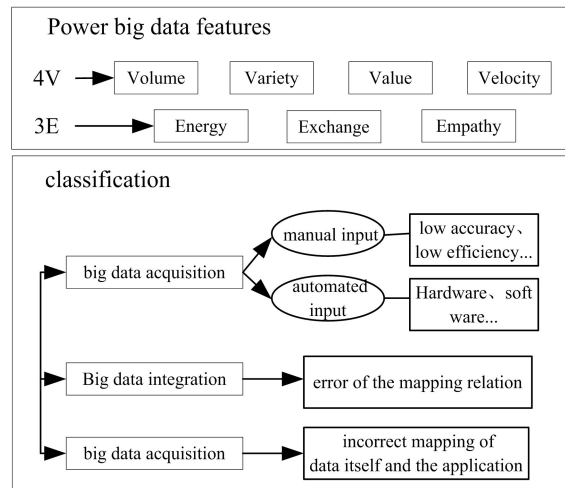


Figure 1. The quality problem classification of power big data

Improve power big data quality solution method

Due to the quality of big data and many aspects related, so in improving big data quality will also involve a lot and related information, and the most common is in the analysis of the quality of the data is involved and related knowledge, namely, whether with domain specific business application knowledge. At present, there are two aspects of data quality research, one is power big data quality assessment and monitoring, the other is power big data quality improvement. As shown in Figure 2.

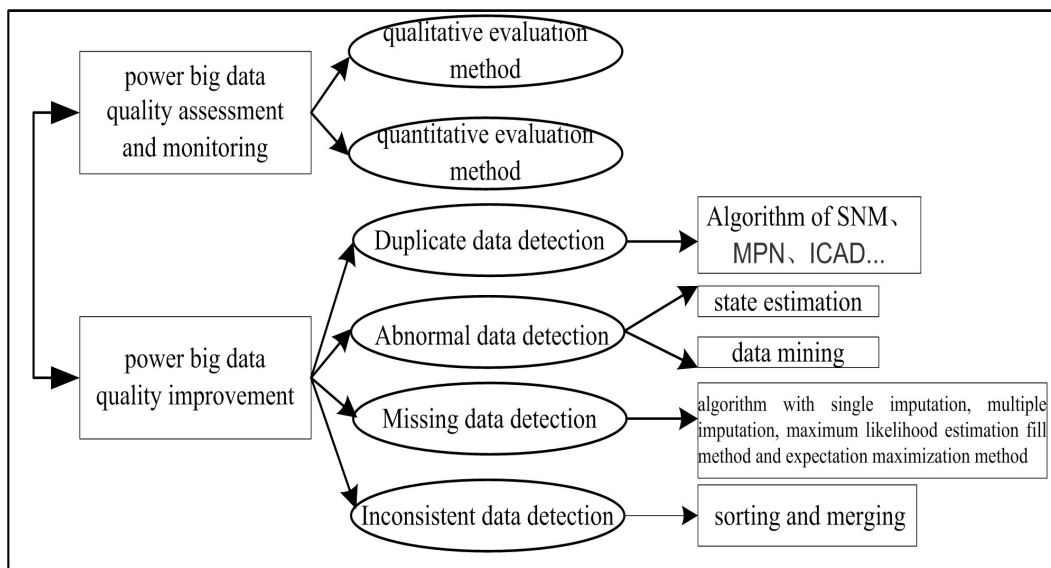


Figure 2. The power big data quality solution method

The Power Big Data Quality Assessment and Monitoring.

Data quality assessment and monitoring the data quality for data analysis, evaluation obtained data results in terms of the quality of the data, the results can be to provide the basis for other applications of data quality. There are two main types of data quality assessment. One is the qualitative evaluation method, the other is the quantitative evaluation method. Qualitative evaluation method is mainly through the qualitative analysis of various data quality dimensions, namely the accuracy, completeness, consistency and timeliness of analysis to assess whether the data meets the requirements. Through such a qualitative analysis to confirm whether or not to meet the requirements of the quality of the data, but such qualitative analysis often don't have the objectivity and reproducibility, it will lead to analysis results may have a larger deviations, and the quantitative evaluation method because it can compensate for these shortcomings become a new research aspect[6].

Power Big Data Quality Improvement

The improvement of the quality of the data is refers to the problems in the technology of data quality problems of data analysis and processing, eliminate data errors and inconsistent, the processing results to meet certain requirements. The improvement of the quality of the data mainly two aspects, are examples of layer and model layer, and the current to improve the quality of data occupy very important technology of data cleaning is mainly layer instances to solve data quality problems, namely through the corresponding detection technology to detect instances of layer number to detect whether there is a such as redundant data, missing data, data quality problems, then according to the corresponding data quality problems take corresponding processing strategy.

Data cleaning technology is widely used in many fields due to its effect on data quality[7]. According to the different types of data processing, data cleaning research has the following aspects.

(1) Duplicate data detection is to detect and process the data that are stored in the. The research mainly focuses on two aspects, one is the duplicate record detection of relational database, the other is the detection of semi-structured data. The algorithm of duplicate data detection has Sorted-Neighborhood method, Multi-Pass Sorted-Neighborhood Method, Improved Clustering Algorithms Using Adjustable Density , and so on[8].

(2) Abnormal data detection. The exception of the data in the generation is often due to two reasons, one is the inherent variability of data, and the other is due to the measurement or execution of errors in the process. So far, the research methods of the detection and identification of power system abnormal data can be summarized as two methods based on state estimation and data mining.

State estimation is to improve the accuracy of the data by using the redundancy of real-time measurement system, and to eliminate the error caused by random interference, and then estimate the running state of the system. The method of detecting and identifying bad data based on state estimation is mainly based on the method of the residual search method, the non-two criteria method, the zero residual method and the estimation method. These methods mainly will weighted residuals or standard error values as feature values, assuming the subject to a certain probability distribution, and in accordance with certain confidence level to determine a threshold value, hypothesis testing is carried out. After finding suspicious data, the new state estimation value is obtained by excluding or reducing its weight.

Analysis method of the data mining are association analysis, sequence pattern analysis, classification analysis and cluster analysis. The detailed algorithm of data mining mainly statistical analysis method, decision tree method, neural network method, covering positive examples and rejection counterexample method, rough set method, concept tree method, genetic algorithm, formula discovery, fuzzy set method and visualization technology. According to the different analysis methods and specific algorithms, based on data mining of bad data detection and identification methods can be divided into based on neural network and fuzzy theory and cluster analysis of the two methods.

(3) Missing data detection. Missing data in reality is often, and this tends to result in a large impact on the results of data processing. So it is very important to improve the quality of data. In detecting to missing data can be used to automatic processing method, the common algorithm with single imputation, multiple imputation, maximum likelihood estimation fill method and expectation maximization method to fill and so on. Different algorithms have different effects, the actual needs to be selected according to the requirements.

(4) Inconsistent data detection. When multiple data sources are integrated in the data collection, the data sources are often stored in the same data, which is easy to produce inconsistent data. To obtain the ideal data from several inconsistent data sources, the common methods are sorting and merging[9].

Conclusions

This paper based on the characteristics of the power of big data to study the quality management method based on the power of big data, including analysis of commonly used data quality evaluation index, classification of data quality problems, on this basis is given to improve the data quality of solutions. Next, it will continue to study more efficient data cleaning methods and how to improve the practical utility of the method in power data.

Acknowledgment

This work is supported by the State Grid Corporation project: (1) Key Technologies for Power System Security and Stability Defense Considering the Risk of Communication and Information Systems. (2) Researches on the privacy prevention theories, models and key technologies in the procession of the data exchange in the environment of the energy-information network. (3) Researches on key technologies and theories of data security in the environment of smart grid big data.

References

- [1]WanTao, Yang Desheng, Sun Fei. Research and application of power big data specialization analysis technology[J], China Management Informationization, 2014, (20),31-33.
- [2]Zhang Pei, Yang Huafei, Xu Yuanbin, Power Big Data and Its Application Scenarios in Power Grid[J], Proceedings of the CSEE, 2014, (z1),85-92.
- [3]Chen Chao, Zhang Shunshi, Shang Shouwei, Sun Fei, Research on power industry data application based on big data[J],Modern Electronics Technique, 2013, (24),8-14.
- [4]zhao Yunshan, liu huanhuan. Research on Big Data Technique Application in Electricity Power Industry[C] 2013 Electric power industry information technology annual meeting, 597-598.
- [5] Chen Chao. Research for electric power big data quality evaluation model and dynamic exploration technology [J]. Modern Electronics Technique 2014(37), 153-155.
- [6] Huang Hui, Zhu Qiliang. Development analysis and Prospect of data quality control in Smart Grid [J]. Science & technology information. 2012(07) ,92-93.
- [7]Chen Mengjie. Research and application of data quality management and data cleaning technology [D], 2013.
- [8] Yang Xi, Tan Jiancong, Zhang Jun, Application of data quality management in electric power production information system[J], Electronics World, 2013(22),234-235.
- [9] Zhang Lan. Analysis of Methods to Improve the Authenticity and Accuracy of Electric Power Statistical Data[J], Value Engineering,2013, (36),183-184.