ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 143

2nd International Conference on Applied Mathematics, Modelling and Statistics Application (AMMSA 2018)

Quality Analysis Method Design for Data from Power Consumption Information Collection System

Sen Pan*, Junfeng Qiao and Jing Jiang

Global Energy Interconnection Research Institute Co.Ltd, Nanjing 210003, China

*Corresponding author

***Abstract*—In the era of big data, the data contains enormous value, and data quality is an important prerequisite for obtaining data value through data analysis and mining. Starting from the definition of data quality, the article first introduced the importance and main indicators of data quality analysis, then analyzed the sources and main characteristics of data collected by electricity consumption in detail. It proposed a quality analysis method based on big data statistical analysis and big data mining analysis, and detailed information were introduced in terms of key technologies, analysis procedures, and indicator design. Finally, an application verification and a simple summary of quality analysis method were given.**

Keywords*—data quality analysis; big data; data from power consumption information collection system; data statistics; data mining

I. INTRODUCTION

Data is the result of facts or observations. It is a logical induction of the information contained in objective things and is the rawest material used to represent objective things. As a carrier and expression of information, data is an identifiable number or symbol, such as numbers, letters, text, images, sounds, images, and so on. In ISO 9000, the definition of quality refers to the level of which a set of inherent characteristics meets the requirements. Data quality refers to the level or characteristics of the information delivered to the user's needs. With the increasing popularity of information technology applications, data quality issues in information systems have received more and more attention, especially in the fields of economy, management, and computers. As the carrier of information, data is the basic element which affects the correctness of the decision. Poor data quality will lead to inaccurate or even wrong decisions, which may cause huge losses to the organization. If the quality of data cannot be guaranteed, the implementation of other processes will not achieve the expected results at all. Therefore, it is necessary to perform quality analysis on the data so as to provide support for follow-up applications.

Because of the crucial effect of data quality, many researchers have conducted a series of studies around data quality analysis and have achieved some results. At the same time, the big data brings many challenges for data quality analysis, and further increases the difficulty of data quality analysis, and also brings new research opportunities to data quality analysis.

Combining the massive features of data collection with electricity information, this paper proposes a data quality analysis method which combines big data statistical analysis and big data mining analysis based on big data technology to improve the efficiency and increase indicator of data quality analysis.

II. DATA FROM POWER USER ELECTRIC ENERGY DATA ACQUISITION SYSTEM

The power user electric energy data acquire system is a system which is used to collecting, processing, and real-time monitoring the power user's power consumption information[1]. It is located at the end of the power grid. It is intended for the majority of power consumers and covers all High and low pressure users and metering gates, to achieve automatic collection of electricity information, anomaly measurement monitoring, power quality monitoring, electricity analysis and management, related information release, distributed energy monitoring, intelligent power equipment information exchange and other functions.

The data from the power user electric energy data acquire system is extremely large, and it includes various users' voltage, current, energy, electricity, and other electricity consumption data and related basic data. It support for the analysis and decision for all aspects of business management, and provide information foundation for intelligent two-way interactive services. Because the power user electric energy data acquire system is a complex and huge system, so it may encounter a variety of unpredictable abnormal conditions during the operation process, such as the abnormality of data collection, abnormality of the acquisition terminal, abnormality of channel, and abnormality of the main station. These system abnormalities will create abnormal data which have quality problems in the data collection data, and will adversely affect the follow-up data analysis[2]. In order to assess the impact of data on follow-up analysis work, quality analysis of the data is required.

The quality analysis of data from power user electric energy data acquire system mainly analyzes the extent to which the data meets the needs of follow-up analysis and decision-making processes. It is usually measured by some indicators such as accuracy, completeness, validity, consistency, and timeliness. Accuracy refers to the extent to which a given set of data values corresponds to a set of corresponding correct values. Completeness refers to the extent to which data remains intact. Validity refers to the extent to which the data meets a certain

range of data. Consistency refers to data that can be expressed by the same format. Timeliness refers to the degree of satisfaction of the application of the time characteristics of the data. Taken together, the quality analysis is to evaluate whether the collected data is consistent with the characteristics of the objective entity, whether there are missing records or missing fields, whether the values of the same attribute of the same entity are consistent in different systems or data sets, whether the system meets the system requirements, Whether the data satisfies defined conditions or a certain range of values, and whether the data is stable and within its validity period.

III. Quality Analysis and Big Data Technology

A. Data Statistics

Data statistics refer to the research activities that are conducted from a combination of quantitative and qualitative methods using statistical methods and knowledge related to the objects of analysis. It is based on the data collection work through analysis to achieve a more profound understanding of the research object. The using of statistical methods, and the combination of quantitative and qualitative analysis were an important feature of statistical analysis.

Data statistics are an important method for conducting scientific research. Through the number to reveal the quantitative characteristics of things in a particular time, in order to quantitatively or qualitatively analyze things, so as to make the right decisions. The combination of data statistics and data quality analysis can quantify and demonstrate the quality of data through various forms of statistical methods to achieve the purpose of data quality analysis.

As a data statistics engine, impala is an open source, native analytic database for Hadoop. Impala can raise the bar for SQL query performance on Hadoop while retaining a familiar user experience. With Impala, we can query data, whether stored in the hadoop distributed file system(HDFS) or HBase, including SELECT, JOIN, and aggregate functions in real time. Furthermore, Impala uses the same metadata, SQL syntax, ODBC driver, and user interface as Hive, which provides a familiar and unified platform for batch-oriented or real-time queries[3].

B. Data Mining

Data mining refers to the process which extracts effective, novel, potentially useful, and ultimately understandable information and knowledge from a large number of uncertain, incomplete, obscure, and random, actual application data stored in databases, data warehouses, or other repositories[4]. Based on data mining, a variety of data mining technologies are applied to the data quality analysis process. Its purpose is to discover and quantify the data quality problems existing in massive data, also known as data quality mining.

Data quality analysis based on data mining technology has many advantages. It can solve many types of data quality problems, improve the scope of data quality analysis, and enhance the objectivity, accuracy, and versatility of the analysis results. Based on the mining analysis of big data technology, the mass data storage platform, parallel high-speed computing, columnar storage database, parallel mining algorithm, and NoSQL database engine are fully utilized to achieve efficient and rapid data mining and analysis of the entire sample data in order to provide technical support for quality analysis of massive data.

Spark MLlib is Spark's scalable machine learning library which is frequently used for data mining. Its goal is to make practical machine learning scalable and easy. At a high level, it provides tools such as[5]:

- ML Algorithms: common learning algorithms such as classification, regression, clustering, and collaborative filtering.
- Featurization: feature extraction, transformation, dimensionality reduction, and selection.
- Pipelines: tools for constructing, evaluating, and tuning ML Pipelines.
- Persistence: saving and load algorithms, models, and Pipelines.
- Utilities: linear algebra, statistics, data handling, etc.

K-means is one of the most commonly used clustering algorithms that clusters the data points into a predefined number of clusters. The spark MLlib implementation includes a parallelized variant of the k-means method.

C. Data quality analysis

Data quality analysis uses a quality analysis methods by the combination of query statistics and data mining[6,7]. Based on the quality analysis of mathematical statistics, through the Impala high speed query engine and the distributed computing framework, the advantages of Impala which is based on memory are used to increase the analysis speed. The quality analysis of data mining based on Spark MLlib scalable machine learning library can achieve the parallelization of the mining algorithm K-means. This simplifies the data mining process and greatly improves the speed and efficiency of data mining.

IV. Quality Analysis Method Design and Application

A. Quality Analysis Method Design

The quality analysis method of data from power user electric energy data acquire system is based on Hadoop's big data cluster such as the hadoop distributed file system, Impala, HBase, Spark MLlib and so on. This method can effectively and quickly realize the quality analysis of the data combined with big data statistics and data mining methods. The analysis method has achieved the quality analysis of data through the steps of data acquisition, preprocessing, statistical analysis and in-depth mining, etc. The main quality analysis process includes three stages: data integration, data analysis, and data displaying.

The data preparation part is mainly responsible for the acquisition of electricity using data. In the process of data preparation, the hadoop distributed file system of the Big Data Platform is used as a storage medium for massive data, and a

variety of data collection methods and tools are used to quickly and efficiently prepare the large amounts of power consumption data. At present, the data from power user electric energy data acquire system is mainly stored in Oracle data and big data platforms. For these data which is in the form of relational databases, the data exchange tools of sqoop are used to import related data tables into distributed storage systems and stored as text; for data stored on big data platforms, copy the data directly to a new distributed storage system. The quality analysis method proposed in this paper also involves some auxiliary data, such as classification standards, coding standards, etc. These data exist in the form of text, and are directly uploaded to the distributed storage system using the FTP tool.

The data integration part mainly completes the acquisition of data and the establishment of wide tables. The data of wide tables comes from power customer table, power meter table, metering point table, power energy table, power voltage table, current consumption table and history power table which have been imported from power user electric energy data acquire system. Through the correlation relationship, one wide table based on electricity customer information could be established, including energy meter information and measurement point information, and were stored in the HBase. Another one wide table based on measurement point information also could be established, including the electricity load, real-time acquisition information such as voltage and current, and historical electricity consumption information, which were stored in the HBase At the same time, the impala wide tables mapped to HBase are created[8].

The data analysis part is used to complete the quality analysis function[9]. This part adopts the impala's memory calculation, and the speed is greatly improved. Based on the data analysis quality analysis method of Spark MLlib, it achieved the parallelization of common mining algorithms which greatly improves the speed and efficiency of data mining. The indicators involved in the data quality analysis method of this article mainly include: consistency, completeness, accuracy rate, and effective rate. For consistency and completeness, it is calculated through query statistics by impala. For accuracy rate and effective rate, it is done through data mining that the normal distribution of the data can be obtained by the clustering algorithm.

After the detailed design and analysis to quality analysis method above, four calculation formulas were given to calculate quality analysis indicators and the format of the result is a percentage, which were shown as:

- Conformance calculation formula:

  "consistent data records/total records * 100%".

- Integrity calculation formula:

  "Non-null records / total records * 100%".

- Accuracy rate calculation formula:

  "records with normal changes / total records * 100%".

- effective rate calculation formula:

  "records with normal values/ total records * 100%".

The output part of data display is giving a summary and display output of the quality analysis results. After the completion of data quality analysis, all the results are summarized and displayed in a variety of visualized ways. It provides many visualization tools, so that each of the quality indicators can use a large number of graphical visualization elements to display.

B. Quality Analysis Method Application

In order to verify the quality analysis method proposed in this paper, the quality of curve data from a month which comes from an electricity company was analyzed and verified from the following aspects: consistency, completeness, accuracy rate and effective rate.

1) Consistency

This is to examine the daily trend of curve data. The record number of daily measurement point voltage curve table, daily measurement point current curve table, and daily measurement point power curve table were analyzed with the record number of electric energy meter, as shown in FIGS. 1, 2 and 3.

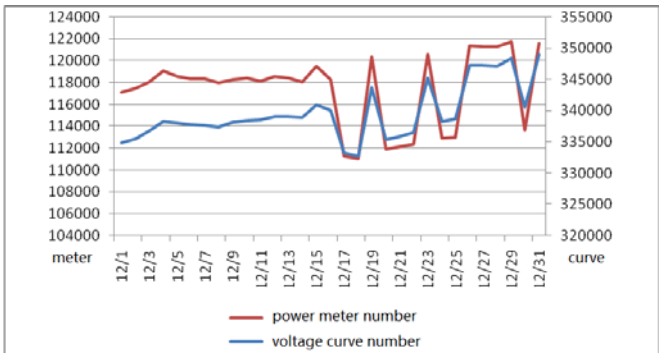

FIGURE I. DAILY VARIATION OF VOLTAGE CURVE DATA

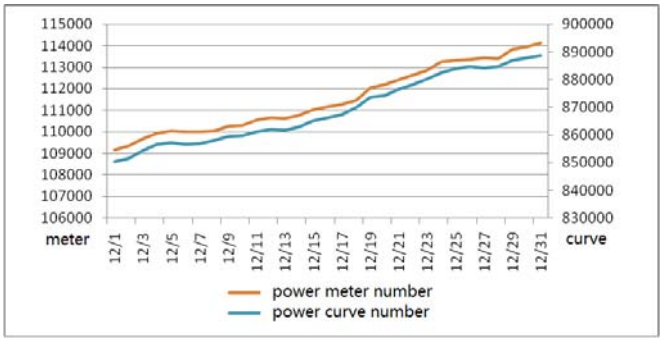

FIGURE II. DAILY VARIATION OF CURRENT CURVE DATA

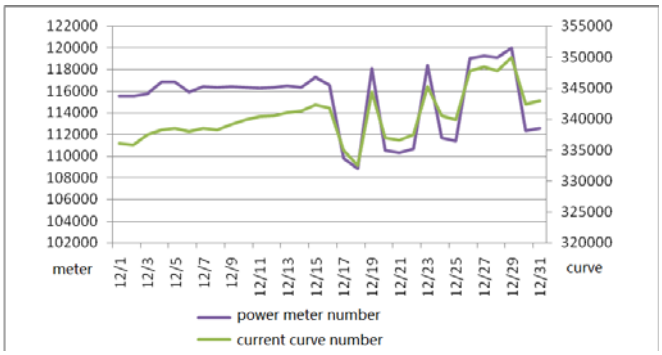

FIGURE III. DAILY VARIATION OF POWER CURVE DATA

From the analysis results above: (1) The daily variation of the voltage curve is basically consistent with the electric energy meter; (2) The current curve appears inconsistent between December 8 and December 15, and the number of electric meters has not seen a substantial increase. However, the number of curves has increased significantly; (3) the growth curve of the power curve has been inconsistent during the period from December 23 to 27.

2) Integrity

Check the integrity of the data items in the curve data table. The analysis results are the ratio of the successful data items collected in the three tables of the daily measurement point voltage curve table, the daily measurement point current curve table, and the daily measurement point power curve table. The data item completeness rates of voltage curve, current curve and power curve were 81.95%, 64.52% and 80.68%, respectively, and the overall completeness rate of the curve data was 74.74%.

3) The overall accuracy rate of the curve data

Examine the accuracy of the data items in the curve data table, and analyze the proportion of the numerical mutation data items in the curve data through the data calculation method. Analyze the data of the 96 points collected from the power curve of the measurement point on the sampling day, and calculate the ratio of 3.73% for the data with numerical mutation. The analysis result shows that the overall accuracy rate of the curve data is 96.27%.

4) The overall effective rate of the curve data

Analyze the valid data in all the curve data, that is, realize the normal collection and non-glitch data items, and calculate the ratio by comparing valid data items to the overall data items. The analysis result shows that the data item effective rate is 80.42%.

V. CONCLUSION

This paper analyzes the characteristics of data from power user electric energy data acquire system, studies the key technologies and methods of big data analysis, and proposes a quality method by statistical methods and mining analysis. This method improves the efficiency of data quality analysis, enriches the indicators of data quality analysis, and solves the problem of the traditional quality analysis methods which meets trouble in mass data such as performance issues. However, the amount of data verified by the examples in this paper is relatively small, so the next step is to conduct data quality analysis with larger data volume to test the accuracy and applicability of this method, and to improve the data quality analysis method based on the analysis results.

REFERENCES

[1] HU Jiangyi,ZHU Enguo,DU Xingang,et al.Application status and development trend of power consumption information collection system[J].Automation of Electric Power Systems,2014,38(2):131-135.

[2] QIAN Lijun,LI Xinjia.Strategy of the data checking and the exception reason analysis in the information collection system[J].Power Demand Side Management,2013,15(1):45-47.

[3] http://impala.apache.org.

[4] WANG Guanghong,JIANG Ping. Survey of Data Mining[J].Journal of Tongji University(natural science),2004,32(2) :246-252.

[5] http://spark.apache.org.

[6] ZONG Wei,WU Feng. The Challenge of Data Quality in the Big Data Age[J].Journal of Xi'an Jiaotong University (Social Sciences),2013,33(5):38-43.

[7] CHEN Chao.Research for electric power big data quality evaluation model and dynamic exploration technology[J].Modern Electronics Technique,2014(4):153-155.

[8] ZHANG Dongxia, WANG Jiye, LIU Keyan, et al. Application of big data technologies in power distribution and utilization system[J], Distribution & Utilization, 2015,32(8): 6-11.

[9] CHEN Jirong,LE Jiajin.Reviewing the big data solution based on Hadoop ecosystem[J]. Computer Engineering and Science, 2013,35(10):25-35.