Research on Data Cleaning Technology of Distribution Electrical Communication Network

Qin XiaoYang[1, a, *], Liu Yan[1], Wang Lei[1], Wu Lijie[1]

[1]State Grid Henan Information & Telecommunication Company, Zhengzhou, 450000, China

Keywords: Distribution Electrical Communication Network, Big Data Analysis, Data Cleaning Technology

Abstract: After the distribution electrical communication network information construction in recent years, a large amount of raw data has been accumulated, including equipment alarm, performance and configuration data, operation and maintenance data and professional management data. The smart device resource can obtain relevant data through the manufacturer network management interface to ensure the accuracy, reliability and integrity of the smart device resource data; but the dummy resource (dummy resource device and dummy resource connection) is an integral part of big data analysis ,"Dumb resources" do not open, big data analysis is difficult. This paper studies the key technologies of data cleaning in the communication network to realize the data and related data cleaning of the terminal communication access network, improve the accuracy, reliability and integrity of the data, improve the quality of the data collection, and meet the needs of data analysis.

1. Introduction

The distribution link is the key link of the whole smart grid to connect the users with the grid. It delivers reliable electricity directly to urban and rural users. To realize and guarantee reliable power supply, power supply quality and two-way interaction to users. Power distribution network is an important part of power distribution intelligence. Its performance and reliability have a decisive influence on the realization of the whole function and operation reliability of distribution automation system. Distribution power communication has the characteristics of large number of communication terminal nodes, scattered communication nodes, short communication distance, small amount of communication data, and great influence of distribution network expansion and urban construction.

Through Collecting multi-vendor configuration, alarm, and performance data on a unified and standard data collection platform ,combining the related dummy resource data, studying the big data acquisition and cleaning management, evaluating the correlation data of the terminal communication access network, it can achieve data accuracy, reliability, and integrity, and transform “incomplete data” into data that meets data quality requirements or application requirements.

2. Power Distribution Network Operation Management Data Acquisition Technology

In the heterogeneous power communication network, including the terminal communication access network[1], the optical cable network, the optical transmission network and the data network, the collection technology is mainly divided into two categories according to the characteristics of different communication data: the first type is the network integrated management system, the device network management system and the device interface protocol obtain multi-operation data about resources, configuration, performance, and faults; the second type is to deploy the probe from the bottom to the link layer, the network layer, the transport layer, the application layer, and the like. The business is fully monitored to obtain network traffic data which will have T-level data volume every day. In the face of growing big data, using data cleaning technology to complete operations such as data conversion, de-duplication, and lack of value, there must be a set of erroneous data processing strategies for erroneous data to improve data quality

DOI: 10.25236/icbsis.2020.041

through various processing methods and provide relatively reliable and reasonable basic data for the subsequent development of data resource.

The multi-operation data mainly refers to data related to resources, configuration, performance, and faults generated during network operation in AMS, TMS, EMS, and NE, and is obtained through data sharing mode and interface protocol [2].

With the continuous deepening of the power enterprise information construction, more and more applications need to access heterogeneous data sources. The problems to be solved in the integration of heterogeneous data sources are complicated. How to use data cleaning technology to eliminate dirty data and ensure the quality of integrated data is an important link. Heterogeneous data source integration is a complex process of importing data from various business processing systems into a comprehensive target database. It needs to process data from multiple business data sources. These data sources may be on different hardware and operating systems. There are large differences in coding, naming, data types, semantics, etc., so how to target Loading these large amounts and types of data in the database has become a key issue for data integration.

3. Data Cleaning Technology of Distribution and Communication Network

3.1 Overview of Data Cleaning Technology

Data cleaning is the last procedure to discover and correct identifiable errors in data files [3], including checking data consistency, handling invalid values, and missing values. Consistency check is based on the reasonable value range and mutual relationship of each variable to check whether the data meets the requirements, and find data that is beyond the normal range, logically unreasonable, or conflicting. The principle of data cleaning is to use related technologies such as mathematical statistics, data mining or predefined cleaning rules to convert dirty data into data that meets data quality requirements. Data cleaning is different from the questionnaire review. Data cleaning after input is generally done by computer instead of manually. The task of data cleaning is to filter the data that does not meet the requirements, and submit the result of the filtering to the competent department of the business to confirm whether it is filtered or corrected by the business unit before extraction. Data that do not meet the requirements are mainly three categories: missing data, erroneous data, and duplicate data [4].

1) Incomplete data: This type of data is mainly due to missing information. For this type of incomplete data, they should be filtered out, and the missing content should be written to different files and submitted to the customer, and they must be completed within the specified time, and then written to the data warehouse after completion.

2) Wrong data: The reason for this type of error is that the business system is not sound enough. After receiving the input, it is not directly judged to be directly written into the back-end database. This kind of error needs to be selected from the business system database by SQL and handed over to the business. The competent department requires a deadline for amendment, and then withdraws after the amendment.

3) Duplicate data: For this type of data, especially in the dimension table, this situation will occur. Export all the fields of the duplicate data record for customer confirmation and collation.

3.2 Data Cleaning Technology

In general, data cleansing is the process of condensing the database to remove duplicate records and converting the remaining parts into a standard acceptable format. The standard model of data cleansing is to input data to the data cleansing processor, "clean" the data through a series of steps, and then output the cleaned data in a desired format. Data cleanup deals with the problems of missing values, out-of-bounds values, inconsistent codes, duplicate data, etc. from the aspects of data accuracy, completeness, consistency, uniqueness, timeliness, and effectiveness. Data cleaning is an iterative process that cannot be completed in a few days. Problems are found and solved constantly. Data cleaning technology is divided into two parts: repeated record cleaning and noise data elimination [5].

1) Duplicate record cleaning: In the process of constructing a data warehouse, you need to import a large amount of data from various data sources. Ideally, for an entity in the real world, there should be only one record in the database or data warehouse. However, when integrating multiple data sources represented by heterogeneous information, there may be various problems such as data input errors, differences in format and spelling, so that multiple records identifying the same entity cannot be correctly identified. Entities that logically point to the same real world may have multiple different representations in the data warehouse, the same entity object may correspond to multiple records. Duplicate logging can lead to erroneous mining patterns, so it is necessary to remove duplicate records from the dataset to improve the accuracy and speed of subsequent mining. Each duplicate record detection method needs to determine if two or more instances represent the same entity [6]. An effective detection method is to compare each instance with other instances to find duplicate instances. However, although this method works best, it is not efficient, time-consuming, and laborious, and is generally not used in reality.

In order to detect and eliminate duplicate records from the data set, the first question is how to determine whether the two records are duplicates. This requires comparing the corresponding attributes of the records, calculating the similarity, and then performing weighted averaging according to the weight of the attributes to obtain the similarity of the records. If the two record similarities exceed a certain threshold, then the two records are considered to be matched; otherwise, they are considered to be records pointing to different entities [7].

The sort-merge method is a standard way to detect completely duplicate records in a database [8]. Its basic idea is to sort the data set first and then compare the adjacent records to be equal. The more common algorithm currently used is the basic neighbor sorting algorithm, which mainly includes the following three steps.

a) Generate keywords: Generate a keyword for each instance by extracting the value of the relevant attribute in the data set.

b) Data sorting: Sort the data in the data set according to the keywords generated in the previous step. As much as possible, the potential possible duplicate records are adjusted to an adjacent area, so that the object matching the records can be limited to a certain range for a specific record.

c) Merging: A fixed-size window is sequentially moved on the sorted data set, and each record in the data set is only compared with the records in the window. If the size of the window contains m records, each new entry window record is compared with the m-1 records previously entered into the window to detect duplicate records; then the record that first enters the window slides out of the window, and finally The next record of a record is moved into the window.

As mentioned above, real-world data can't be directly used for data mining, and it needs to be pre-processed to meet the required data before it can be used for mining.

2) Eliminating noise data: Noise data can appear for a variety of reasons. Due to the presence of noise data, the data is not in the specified data domain, which will affect the subsequent mining effects and results. A common method of eliminating noise data is the binning method [9].

The binning method smoothest the data values that need to be processed by reference to the values of the surrounding instances. The data that needs to be processed is distributed to some boxes, and different binning techniques smooth these values differently. The existing binning methods are the equal-depth binning method and the equal-width binning method. The equal-depth binning method divides the data into different bins of the same depth. The specific method is as follows.

•Smooth by box average: This method averages all the values in the box and then uses the average of the boxes to replace all the data in the box.

•Smooth by box boundary: The maximum and minimum values in the box are considered box boundaries, and each value in the box is replaced by the nearest box boundary value.

In the actual distribution network communication system, the data is continuously uploaded to the primary station in the form of data stream, and the scale of the concurrent transmission is huge, and it is difficult for the computer to accurately analyze the entire appearance of the plurality of periodic data. The distributed parallel framework for data anomaly detection and repair can provide real-time data cleaning for high concurrency online. The distributed architecture of data cleaning with electric communication network is divided into data acquisition and storage module, data preparation module and data cleaning module. The specific functions are as follows.

a) Data collection and storage: collecting from a remote load terminal through a certain communication protocol or using a SQL statement to poll the various data source system databases of the distribution network. Data analysis, differentiation processing and time stamping are performed on the obtained raw data; the massive real-time database is used to cache massive concurrent data, and the flow control strategy is added to optimize the processing delay; and the data input bandwidth is restricted by communication with the data cleaning module. The module is also responsible for data archiving and the provision of data interfaces to advanced application analysis systems.

b) Data preparation: multiple cycle load data is read into the sliding cleaning window, and the sliding window is updated in time to repair the completed "clean" data. In order to avoid generality, this study proposes a "stepped" distributed sharding strategy in a distributed framework. The "ladder" sharding strategy improves the coverage of the correct data in the input sample set of each cluster node by adjusting the slant angle θ and the vertical step h. The relationship between the amount of data ESIZE involved in the calculation and the slice angle and the vertical step size is

$$E_{size} = \frac{hl}{\tan\theta} \qquad (1)$$

c) Data cleaning: transferring the load data after fragmentation to parallel computing in each distributed node, realizing data identification based on the real-time identification algorithm of abnormal load data described in this paper, then implementing data merging on a single computer, using collaborative filtering The recommended algorithm implements the abnormal load data correction, and finally updates the abnormal value in the data preparation module cleaning window and sends the "clean" data to the database archive.

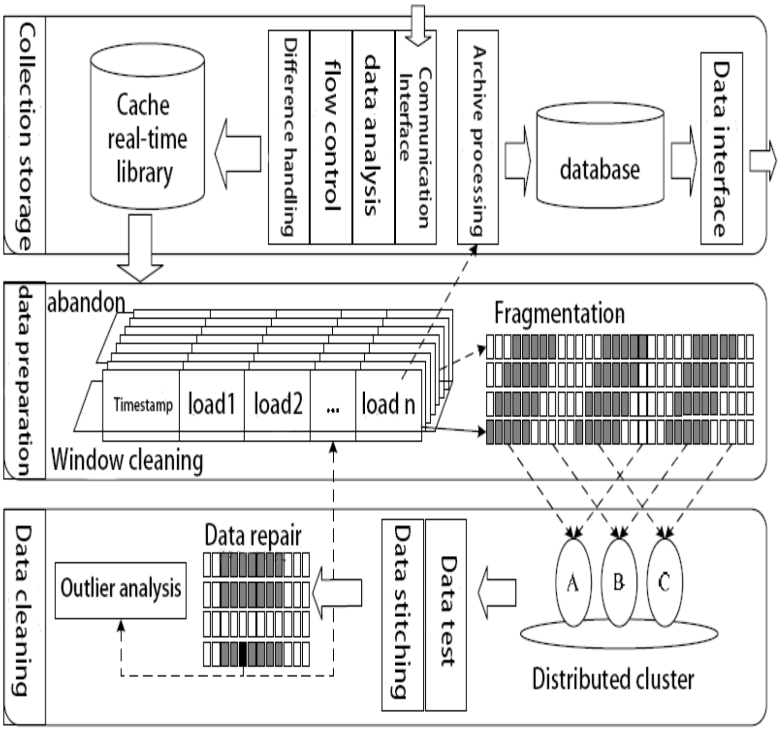

Fig.1. Distribution network data cleaning distributed architecture

4. Conclusion

Data has been widely regarded as a resource. It is a resource that we can use and obtain value and knowledge from. Analyze and mine data resources so that we can make timely, cost-saving, high-quality decisions and conclusions The purpose of data cleaning is to extract data that is valuable and meaningful for solving problems from a large amount of data that is complex, chaotic, and difficult to understand. Save the truly valuable and organized data and do the data analysis later to reduce the analysis obstacles.

In order to further adapt to the needs of "big operation" management reform, improve the intelligent management of distribution communication network, and improve the network support capability, this paper effectively avoids usefulness through reasonable data cleaning methods based on the key technology of data acquisition with electric communication network. It can effectively reflect the dynamic changes of the original data, adapt to the characteristics of the state data of the grid equipment, realize the data of the terminal communication access network and the associated data cleaning, and provide an important basis for the advanced application of the later data analysis.

References

[1]Zhao Teng, Zhang Yan, Zhang Dongxia. Big data application technology and prospect analysis of smart distribution network [J]. Grid technology, 2014,38 (12) : 3305-3312.

[2]Liu Keyan, Sheng Wanxing, Zhang Dongxia, et al. Research on big data application demand and scenario analysis of smart distribution network [J]. Chinese journal of electrical engineering, 2010,35 (2) : 287-293.

[3]Zhang Dongxia, Miao Xin,Liu Liping, et al. Research on the development of smart grid big data technology [J]. Chinese journal of electrical engineering, 2015,35 (1) : 2-12.

[4]Song Yaqi, Zhou Guoliang, Zhu Yongli. Status quo and challenges of smart grid big data processing technology [J]. Grid technology. 2013,37 (4) : 927-935.

[5]Mao Lifan, Yao Jiangang, Jin Yongshun. Abnormal data identification and missing data processing for medium - and long-term load forecasting [J]. Grid technology, 2010,34 (7) : 148-153.

[6]Zhang Le. Data restoration and correction in annual load forecasting [J]. Grid technology, 2007,31 (S1) : 233-234.

[7]Wang Yanping, Le Chunxia. Data preprocessing technology for power system load modeling [J]. Grid technology, 2007,31 (S2) : 292-294.

[8]Zhang Pei, Wu Xiaoyu, He Jinghan. Application overview of big data technology in active distribution network [J]. Electric power construction, 2015,36 (1) : 52-59.

[9]Yan Yingjie, Sheng Geao . Abnormal detection method of state data of power transmission and transformation equipment based on big data analysis [J]. Chinese journal of electrical engineering, 2015,35 (1) : 52-59.