A Power Big Data Processing Method for Heterogeneous Data Sources

Du Hongjun¹, Li Wei¹, Wang Ou¹, Zhang Wenjie*¹

¹ State Grid Liaoning Electric Power Co. LTD Information & Telecommunication Branch Company, Shenyang Liaoning 110000, China
*Corresponding author’s e-mail: sg_eiconf@126.com

Abstract. With the advent of the era of electric power big data, the volume of electric power data will grow at an unprecedented speed. In this paper, we propose a construction scheme of electric power big data service platform. The platform is positioned as a data sharing, analysis and application platform. Data acquisition, storage, preprocessing, visualization and other aspects can be realized. This paper analyzes the demand of platform construction and designs the platform architecture. At the same time, we introduce the implementation way of big data platform from the aspects of overall platform architecture, functional architecture and technical architecture. And the power big data mining algorithm based on Apriori algorithm is introduced. It can be seen that this platform can provide a reliable analysis and mining platform for data mining in power industry.

1. Introduction
With the advent of the era of electric power big data, the volume of electric power data will grow at an unprecedented speed. In the application of power big data, the correlation analysis of data such as energy and weather inside and outside the industry will directly lead to the increase of power data types. The real-time processing characteristics of power big data make it more demanding for the speed of data processing and analysis. At this time, the traditional data mining method will face great difficulties, and how to extract valuable knowledge from a large number of original data becomes a research difficulty.

Based on the above reasons, this paper proposes a construction scheme of electric power big data service platform. The platform is positioned as a data sharing, analysis and application platform. Data acquisition, storage, preprocessing, visualization and other aspects can be realized. This paper analyzes the demand of platform construction and designs the platform architecture. At the same time, we introduce the implementation way of big data platform from the aspects of overall platform architecture, functional architecture and technical architecture. And the power big data mining algorithm based on Apriori algorithm is introduced. It can be seen that this platform can provide a reliable analysis and mining platform for data mining in power industry.

2. The source of power big data
This chapter analyzes the characteristics of power big data. The requirements of bad data identification and heterogeneous data fusion are mainly discussed. The typical data mining methods and Hadoop big data processing architecture are analyzed. Power big data covers the whole process of power system distribution and distribution. From the subsystems of the power system. It has the characteristics of large amount of data, rich data types, low value density, fast processing speed and high authenticity.
Figure 1. The source of power big data.

The power large data source is shown in figure 1. It can be seen that power big data has the characteristics of large quantity and rich types. Power big data comes from different application systems, including SCADA, EMS, WAMS, PMIS, GIS, OMS, etc. However, the independent development of each application system makes it impossible to interact with each other, and the heterogeneous data is not conducive to the unified management of data. Therefore, multi-source heterogeneous data fusion technology is very necessary for the analysis and processing of power big data. In order to realize the fusion of multi-source heterogeneous data in power system, it is necessary to clean the data. Data integrity, consistency, and correctness are critical to data analysis. Especially when the data is missing, it will destroy the accuracy and integrity of the data.

3. Data mining method of power big data

Data mining is the key to transform data into knowledge. Generally, data mining is defined as revealing unknown and potential value information in mass data through a certain algorithm, and then transforming it into knowledge. The process of concluding a general law. Different from the traditional statistical analysis of samples, data mining focuses on the overall analysis of big data. The complete data mining process is shown in figure 2. The specific steps are:

1) Identify data mining objectives

The determination of the target is the first link in the data mining process, and it is also a link with direction guidance function. Business orientation is defined by defining goals, i.e. determining what data and knowledge is meaningful and valuable.

2) Data sorting

After determining the mining target, it is necessary to understand the background information and store the mass data collected in the corresponding database, so as to make preparations for the data mining link.

3) Data mining

According to the characteristics of the data and the requirements of the user or the system, the appropriate data mining algorithm is determined, and the data prediction and correlation are realized.

4) Explain the evaluation

If the data mining results are redundant, missing, or do not meet user requirements, they need to be deleted. Meanwhile, repeat the process from (1) to (3). If the data mining results meet the validation criteria or can meet the needs of users, then the results need to be parsed and expressed in natural language.
4. Data mining process based on Apriori algorithm

Cluster mining is the most typical data mining method to study the possible similarity between data objects and classify them. Apriori algorithm is a typical association rule algorithm, which realizes data mining by generating layer by layer tests.

Apriori algorithm is scanned layer by layer, and $k$ item set is obtained by $k-1$ item set search. Scan the transaction database for the first time get frequent item sets in item 1 for $L_1$. The first $k$ before scanning, first using the frequent item sets, generate $k$ candidate item set, remember to $C_k$. Then traverse the $C_k$ elements in support, through the comparison with $\text{minsup}$ frequent item sets. Repeat the process until the candidate item set is empty when the algorithm stops. The basic flow of Apriori algorithm is shown in figure 3.

Furthermore, Apriori algorithm is implemented based on Hadoop big data processing architecture. At this point, users developing distributed programs can eliminate the process of understanding the low-level details of distribution. Hadoop implements a Hadoop Distributed File System (HDFS). HDFS has excellent fault tolerance, high throughput, and the ability to stream data from file systems, making it suitable for applications with very large data sets. The core design of Hadoop's framework is HDFS and MapReduce. HDFS provides storage for massive amounts of data, while MapReduce provides computing.
5. Performance test and analysis

In order to demonstrate the technical superiority of Apriori algorithm designed in this paper in the field of big data mining. This section compares the advantages of Apriori algorithm and traditional FP algorithm in power big data mining. The analysis process is based on load data of 1 million power users in an area. At the same time, combined with local climate, environment, major events and other multiple factors, to carry out load law prediction and analysis. A total of 8 typical days were selected for load prediction and analysis. The results are shown in Table 1. It can be seen that the Apriori algorithm is more accurate than the traditional FP algorithm in data mining.

| Data sets   | FP algorithm Correct rate (%) | Apriori algorithm Correct rate (%) |
|-------------|--------------------------------|-------------------------------------|
| Group 1     | 91.95                          | 92.44                               |
| Group 2     | 88.70                          | 93.35                               |
| Group 3     | 82.90                          | 94.59                               |
| Group 4     | 88.08                          | 90.19                               |
| Group 5     | 86.46                          | 89.54                               |
| Group 6     | 93.43                          | 94.27                               |
| Group 7     | 84.59                          | 94.22                               |
| Group 8     | 89.71                          | 96.25                               |
| Average value | 88.20                          | 93.15                               |

Figure 3. The basic flow of Apriori algorithm.
6. Conclusion
A power big data processing method for heterogeneous data sources is proposed in this paper. The platform is positioned as a data sharing, analysis and application platform. Data acquisition, storage, preprocessing, visualization and other aspects can be realized. This paper analyzes the demand of platform construction and designs the platform architecture. At the same time, we introduce the implementation way of big data platform from the aspects of overall platform architecture, functional architecture and technical architecture. And the power big data mining algorithm based on Apriori algorithm is introduced. It shows that the platform can provide a reliable analysis and mining platform for data mining in power industry.

References
[1] Wang B, Fang B, Wang Y, et al. Power System Transient Stability Assessment Based on Big Data and the Core Vector Machine[J]. IEEE Transactions on Smart Grid, 2016, 7(5):2887-2894.
[2] Reitenbach G. Big Data and the Industrial Internet Meet the Power Plant[J]. Power, 2016, 160(1):26-31.
[3] Munshi A A, Mohamed A R I. Big data framework for analytics in smart grids[J]. Electric Power Systems Research, 2017, 151:369-380.
[4] De Caro F, Vaccaro A, Villacci D. Spatial and Temporal Wind Power Forecasting by Case-Based Reasoning Using Big-Data[J]. Energies, 2017, 10(2):252-261.
[5] Zhu J, Zhuang E, Fu J, et al. A Framework-Based Approach to Utility Big Data Analytics[J]. IEEE Transactions on Power Systems, 2016, 31(3):2455-2462.