The Application of Power Network Data Mining and Optimization processing based on Distribution network

Allam Maalla*, Chao Jia, Xiaohong Ning

Research Institute of Open Data Commercial Application, School of Information Technology and Engineering, Guangzhou College of Commerce, Guangzhou, China. 511363
allamaalla@yahoo.com

Abstract. Distribution network has a wide range of business, and different users have different needs in business data. Mining the data of interest of distribution network business personnel can process business more efficiently and improve work efficiency. Focusing on data mining model and data mining algorithm, the project needs data processing before and after mining, including deleting repeated data and data compression, accelerating data processing speed and optimizing platform construction. The application platform can push relevant information and reports according to users' demands, reduce users' time in searching grid data, present users' required information more efficiently, and improve the overall working efficiency and level of network personnel.

1. Introduction
With the rapid development of data, data mining technology has been developing continuously since the 1980s. At present, it is mainly used in the fields of finance, communication, transportation, manufacturing and retail, which have massive data. Power industry, as an industry that also produces massive business data, is also a very important direction for the application and development of data mining technology. Currently, there are many data mining products, mainly introduced as follows:

(1) Rapid Miner. Was originally developed in 2001 by a team of artificial intelligence at the university of Dortmund that could conduct data mining experiments consisting of a large number of arbitrarily nested operations. All operations are described in XML files and can easily be created through Rapid Miner’s graphical user interface. Rapid Miner provides up to 500 functions for almost all major data mining functions.

(2) Mine Set. Mine Set is a product jointly launched by SGI and Sandford University in the United States. This is a data mining product for a multitasking system. Its advantages are advanced visual display and embedded with many data mining algorithms and graphical tools, which can bring users the knowledge behind big data directly and timely. The product is easy to operate and can publish the results of data mining through the WEB.

(3) Informatica. Informatica is a commercial software developed by Informatica, founded in 1993, which can support a number of complex enterprise Data integration plans, including enterprise Data integration, Data quality control, master Data management, B2B Data Exchange, application information life cycle management, complex event management, super message and cloud Data integration. Its component Informatica Power Center is used to access and integrate data of almost any business system and any format. It can deliver data in the enterprise at any speed, featuring high performance, high scalability and high availability.
Microsoft SQL Server 2005 is a platform provided by Microsoft to develop intelligent applications that can use these tools throughout the life cycle of data. Therefore, data mining results are no longer limited to the use of a few specialized analysts, but open to the entire organization. The model is extensible, and third parties can add custom algorithms to support specific mining requirements and run data mining algorithms in real time, allowing real-time verification of data mining results. In the field of overseas wind power generation, UK National Grid (NG) carries out real-time analysis on the wind power generation capacity of its wind power plant in Romania through data mining technology and predicts the generation capacity in a certain period in the future. In the Electricity sales market of Australia, the National Electricity Market (NEM) used data mining methods of support vector machines to predict the average power price.

2. Purpose and significance

The construction of enterprise formalization is divided into three levels: the first level is to realize the automation of basic business through the core business system; The second level is to streamline the upstream, middle and downstream management of enterprises through management information system (ERP, CRM, SCM, etc.). The ultimate goal is to realize the scientific realization of strategic decision-making, tactical development and battle plan, implementation, monitoring, analysis and adjustment through decision support system, and lay a solid foundation for the enterprise's profit and risk prevention.

With the advent of the mass data era, the Internet has brought not only unprecedented and unique opportunities, but also great challenges to users and enterprises. How to store data efficiently and safely has become a new problem for us. Because of the obvious defects of the traditional storage methods, they cannot meet our growing demand for big data. First, hardware damage, data loss, aging and damage of hardware storage devices will bring catastrophic damage to the data. There is almost no hope to restore the original data, and it needs to pay a high price. Second, hacking and privacy leakage. Traditional storage technology, with sharing as its basic feature, cannot guarantee the security of data in the transmission process accurately.

3. Limited file and transmission. Due to the limited storage space and insufficient broadband provided by service providers, traditional storage cannot timely store and transmit big data at high speed. The LZ77 algorithm and LZ78 algorithm were proposed by Israeli scientists Jacob Ziv and Abraham Lempel in two papers respectively in 1977 and 1978, and have been playing an important role in the field of data compression for a long time. LZW algorithm evolved on the basis of LZ78 algorithm, which has good effect on data compression, is fast and easy to be accepted, and is a very common lossless compression algorithm. However, the defect of this algorithm is also very obvious. It is neither suitable for the compression of too large file, nor for the compression of too small file, and has obvious fixed word structure. Therefore, we propose an improved LZW algorithm on this basis, which overcomes the above disadvantages and makes it significantly improved in data compression time and compression ratio, providing reference value for subsequent relevant studies. Therefore, the study of this algorithm has important academic value.

With the rapid development of cloud computing and computer hardware technology, especially in the fields of high-definition television, digital broadcasting, bioengineering, defense industry and aerospace, more and more information needs to be processed by computers. This presents a huge challenge to computer performance, especially for hard disk storage and information transmission, both in space and time. To solve this bottleneck, save storage space and effectively improve the transmission speed of information, we need to compress the stored data and ensure that the compressed data is lossless. Therefore, it is valuable to study the data compression technology.

3. Main technical contents

3.1 Demand statistics of network users
Conduct in-depth research and investigation in the power grid, understand the current business content and existing problems of power distribution network, and have a deeper understanding of the demand of power distribution network users.

3.2 Grid data mining model selection

On the basis of understanding the needs of the users of the distribution network, the existing data mining model is classified and analyzed, and the data mining model suitable for the users of the power network is selected to prepare for the next development.

3.3 Set up the data mining platform based on the needs of users of distribution network

The overall system architecture design should meet the adaptability, high efficiency and expansibility under the typical frame of south net:

The achieved data rate at relay node \( k \) in the first time slot and at node \( j \) in the second time slot after using maximum ratio combining are given, respectively by:

\[
 r_r(i,j) = \log_2 \left( 1 + g(i,k)P_d(i,j) \right)
\]

\[
 r_r(i,j) = \log_2 \left( 1 + g(j,k)P_d(i,j) \right)
\]

\[
 r_c(i,j) = \log_2 \left( 1 + g(i,j)P_d(i,j) + g(k,j)P_r(r(1,2)(k,j)) \right)
\]

Note that for relaying to be useful, the achieved rate at relay node \( k \) must be higher than that due to direct transmission. Let the maximum achievable data rate on link \( (i,j) \) be \( r(i,j) \). This achievable rate on each link depends on the transmission power, link gain, and the transmission strategy, the achievable transmission rate on link \( (i,j) \) for different transmission strategies can be written as

\[
 r(i,j) = \begin{cases} 
 r_d(i,j) & \text{for direct transmission} \\
 r_c(i,j) = r_c(i,j) & \text{for cooperative transmission} 
\end{cases}
\]

(2)

On the premise, it realizes the function of index display and data analysis and prediction of distribution network business platform. Furthermore, the system architecture platform is designed to accommodate the growth of data and the smooth extension of the system as user’s increase. The system architecture design is mainly based on the selection of data mining models to determine the appropriate application products (IBM Cognos, Informatica, SAP BW, DB2, Sybase IQ) for the system selection, so that the architecture can meet the requirements and be extended.

4.main technical and difficulties

Design: according to typicality of southern power grid in the southern power grid typicality, on the basis of design requirements, the southern power grid ODS data center BW typical southern power grid and the typical design data warehouse design, form the data layer, data extraction, ODS layer, data warehouse, data mart and front-end display layer data such as hierarchical architecture. The ODS layer mainly serves as the interface of various source data systems and data transformation sharing area. The EDW layer integrates enterprise data and retains massive historical data for future expandable business analysis needs. Data extraction layer includes data extraction conversion load of source system -> ODS, data extraction conversion load of ODS -> EDW and data extraction conversion load of edw -> DM. The front-end presentation layer is mainly based on DM. The principle of universality: the positioning of the data mining system itself should be a part of the enterprise basic system, and its function design and architecture design should be irrelevant to the business. No matter what kind of business data will be organized into this system in the future for data calculation, analysis and demonstration, it can be quickly deployed and realized, and the secondary development when increasing business analysis can be minimized.

5.Construction of the mining platform

The platform should be designed with a design margin and be able to adapt to changes in the future. As far as the actual grid is concerned, there are two predictable changes:
The first is the change of user demand. With the continuous development of enterprises, there will be new requirements for management, and the trend of refinement and flattening will naturally increase the demand for data analysis. From a horizontal perspective, the measurement and indicators concerned by the middle and upper levels of the enterprise will constantly increase or change. From a vertical perspective, the requirements for the information detail degree of data response and the multi-dimensional data slice analysis will gradually increase. Meeting these changes requires the system to be able to adapt quickly.

The second is the use of scene changes. With the acceleration of enterprise system application adjustment and integration, the improvement of future information system engineering will usually make adjustments to its data organization form, which requires the data mining system to be able to adapt to such adjustments.

The principle of high efficiency: every step of the data mining system, there will be a huge amount of data to be processed. From the point of view of satisfying user demand, only the high efficiency of the system can guarantee the availability of the system.

According to the data processing process, the operation monitoring platform can be divided into three parts: the source data layer, the data warehouse layer and the presentation layer. There is a dependency between the three logical layers of the support system. Among them, the presentation layer is responsible for the comprehensive presentation of the data processed by the application service layer, and the two interact through Web services, messages and data integration. The data warehouse layer analyzes and summarizes data in various business support systems located in the source data layer according to business themes, calculates various indicators, forms process examples, and provides analysis results for the presentation layer. The source data layer manages the data of each application and provides business detail data for the data processing layer, as shown in the figure. The overall design scheme of this platform should meet the following three key functional modules: data storage, data mining process, and terminal data presentation.

![Figure 1. Construction of the mining platform](image)

Data warehouse is designed in two layers, namely Operational Data Store and DW (Data Warehouse) /DM (Data Mart) layer. Integrate "three sets of five" business application systems, which provide data support for operation monitoring business through data integration, application integration and interface integration. The data in the source data layer needs to be processed by ETL tools and loaded into the data warehouse layer. According to the deployment requirements of southern power grid and the actual situation of the provincial company, the ETL USES two products and is completed in two steps, one is SAP BW, used to extract data of ERP, Second, Informatica is used to extract data of systems other than ERP and data output of SAP BW.

Although some high coding is wasted, the implementation efficiency of the algorithm is improved, and the compressed data is optimized. For the LZW algorithm, it relies on the method of fast
dictionary lookup, such as using a lot of fast sort and half-search, resulting in a time complexity of $0(n^2)$ and overall code execution is inefficient. Adopting a strategy of substituting hash table for quicksort can effectively improve the searching speed of the dictionary and reduce the time complexity to nearly zero ($n$). However, all of these solutions are proposed under the influence of a single factor without comprehensive consideration, especially in the mass data environment, whether these solutions can still exert maximum efficiency is worthy of further study.

6. Conclusion
The research of this paper is mainly based on the combination of theory and practice. In this paper, data mining model selection is taken as the analysis basis, and the entire research is carried out on the basis of experiments. It includes two parts, namely visual large screen and desktop terminal, and through visual large screen and desktop terminal display carriers, it can display data and information in five parts, including comprehensive monitoring, operation analysis, coordination control, panoramic display and comprehensive management.

Acknowledgments
This research was financially supported by the 2017 Higher Education Teaching Research and Reform Project of Guangdong Province, "The Big Data Talent Cultivation Practice Teaching System of Business Schools" (Project No. 2017SJXGG01), and Characteristic Key Subject of E-commerce Construction Project of Guangzhou College of Commerce Foundation, (Project No.TSZDXX201601).

References
[1] Chum O, Werner T, Matas J. Epipolar Geometry Estimation via RANSAC Benefits from the Oriented Epipolar Constraint[C].International Conference on Pattern Recognition. IEEE, 2004:112-115 Vol.1.
[2] Indelman V, Williams S, Kaess M, et al. Information fusion in navigation systems via factor graph based incremental smoothing[J]. Robotics and Autonomous Systems, 2013, 61(8): 721-738
[3] Maalla A, Zhou C, Guang-Yu W U. The Construction of Heterogeneous Platform of Unified Service System Based on Cloud Computing [J]. 2017(cmee). DOI: 10.12783/dtce/cmee2017/20056.
[4] Forster C, Pizzoli M, Scaramuzza D. SVO: fast semi-direct monocular visual odometry[C] //Proceedings of IEEE International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2014: 15-22.
[5] Zhang F. The Schur Complement and Its Applications [J]. Springer-Verlag, New York, 2005, 4(2): xvi
[6] Mur-Artal R, Montiel J M M, Tardos J D. ORB-SLAM: a versatile and accurate monocular SLAM system [J]. IEEE Transactions on Robotics, 2015, 31(5): 1147-1163.
[7] Zhang Zhigang, Zhao Liqiang, Gao Tianying. Research on training program of computer science professionals for big data applications [J], Tianjin Education. 2017, (1): 32-33.
[8] Zhang Fan. Thinking on the cultivation of computer professional application talents in the direction of mobile internet [J]. Science and Technology Information. 2015, (24): 129-130.
[9] Jang, J.S.R. Machine Learning Toolbox Software. Available online: http://mirlab.org/jang/matlab/toolbox/machineLearning (accessed on 23 December 2014).
[10] Maalla A (2016) Construction of Information Ecosystem on Enterprise Information Portal (EIP). J Electr Electron Syst 5: 205. doi:10.4172/2332-0796.1000205.
[11] He, W.; Wu, K.; Zou, Y.; Ming, Z. WiG: WiFi-based gesture recognition system. In Proceedings of the 24th International Conference on Computer Communication and Networks (ICCCN), Las Vegas, NV, USA, 3–6 August 2015; pp. 1–7.