Research of data mining system

Xu Ruiying

Baicheng Normal College

Keywords: Data mining system; database (KDD); mobile environment

Abstract. Combining the current development situation of data mining system, this paper introduces the centralized and distributed data mining system respectively, focuses on the detailed introduce to the various components of the centralized data mining system and its concrete realization technology, and summarizes the current development situation of the centralized and distributed data mining system at the same time respectively, then puts forward the research direction and development trend of the data mining system: to enhance the visualization and interaction, to improve the scalability, , follow the single standard combined with a specific industry application and support of data mining of the mobile environment. The development of data mining system has been made into a brief summary and outlook.

Data mining, also known as knowledge discovery in database (KDD), is that the process of digging out interesting knowledge from a large amount of data stored in the database, data warehouse or other information. In recent years, to promote the actual application of data mining, many researchers do a lot of research work on the architecture of data mining system. A data mining system with reasonable structure should have the following features:

1) completeness of system function and auxiliary tools;

2) scalability of system;

3) support for multiple data sources;

4) processing capabilities of the large amount of data;

5) good user interface and showing ability of the results. In the current, the data mining system mainly includes centralized and the distributed data mining system, and the specific structure and its various components of each system have a variety of different implementation techniques and methods.

Centralized data mining system

The single database/data mining system is the current data mining application system that has a more mature development, and many commercial data mining application software are based on the structure. From the analysis of the current main data mining system, it can be found that the specific implementation techniques of different products to various functional modules are not the same.

User interface and knowledge presentation layer. In this layer, showing mining results by providing a friendly user interface and using the data visualization technologies can greatly improve the usability of the system. The visualization of data mining is to use visualization technology to find out implicit and useful knowledge from a large number of data set. The visualization of data mining mainly includes the visualization of data, mining process and mining model. The current visualization techniques mainly include the traditional geometry method (such as graph, histogram, scatter plot, pie charts, etc.) SOM network visualization technology, parallel coordinates technique, the visualization technology facing the pixels, etc. The visualization technology based on the SOM network and the parallel coordinate are the two that have more applications, and the principle of them are to display data in the two-dimensional plane through high-dimensional data mapping for two-dimensional data. Such as a visual mining system based on the SOM network VISMiner designed by Wang Jiacai and others, and Liu Kan and others studied the specific application of parallel coordinates technique in data mining system.

Control layer. Control layer is used to control the execution flow of the system, and coordinate the relationship between each feature and their execution order, which mainly includes the analysis of data mining task, and judgment of data involved in mining tasks and data mining algorithm

should be used according to the task analysis result. The Data mining task is generally defined and explained by the data mining language, nowadays, many researchers put forward their own data mining languages, which are class SQL language structurally, such as DMOL language [1], but they didn't realize the standardization of mining language. In March 2000, Microsoft introduced a new Data mining language specification OLE DB for Data Ming, taking a big step towards the data mining language standardization. Amir Netz introduced in detail how to apply OLE DB for DM specification to the data mining system.

Data source layer. In order to improve the consistency and integrity of data, before data mining, at first the data dispersed and stored in the multiple data sources should be integrated into a unified database/data warehouse through the pretreatment operation of data cleaning and data integration. In order to improve the scalability of system, the specific database products the data source using are shielded, and database interface should use ODBC, JDBC, or OLE DB technology, to change the data source. Zhao Zhihong, Qian Weining respectively proposed data mining system framework and their applications based on data warehouse and large-scale database. The database can be integrated to data mining in the system by four forms: no coupling and loose coupling, half of loose coupling and tight coupling. Tight coupling mode is the most ideal, that is, querying data mining optimized into a cycle and retrieval process so as to combine the two through the data mining, so that we can make full use of data processing functions such as query and summary the database, reduce the development burden of the data mining system, and improve the efficiency of system. Rosa Meo put forward a kind of data mining system framework using data mining language Mine Rule to implement tight coupling with database.

Data layer to be mined. The layer provides data set to be mined meeting the requirements of the data mining algorithm for data mining layer, data set to be mined is formed by data preprocessing operations such as data transformation and data specifications of data related to the mining tasks in the data source layer. In addition to mining data based on directly database /in data warehouse, data mining can also be based on the on-line analytical processing (OLAP) to process, called on-line analytical mining (OLAM). Because OLAM can combine the two, and give full play to the advantages of them, so it can make data mining with high efficiency and good interactive. Jia wei Han professor put forward a kind of OLAM system structure framework integrated by OLAP and DM, and developed a data mining system DBMiner based on the structure. Sanjay Goil studied a kind of system architecture integrated by extensible OLAP and data mining based on parallel processing technology.

Mining layer. This layer is the core of the data mining system, and concrete implementation of the layer is directly related to the functionality and scalability of the whole system. Data mining mainly includes the concept/class descriptions, association rules analysis, classification and prediction, clustering analysis, outlier analysis and evolution analyses several types of model of mining. According to various types of model, people put forward a variety of different implementation algorithm, then that a specific data mining system should include what type of pattern mining algorithms is depended on the system developed purpose and its oriented specific application field. In order to improve the scalability of the system, many systems use component technology to implement the data mining algorithm and their managements. In the current, the more mature component technologies mainly include COM/DCOM, EJB/Java RMI and CORBA/IIOP. Component refers to the composition module which can be clearly identified and has certain functions in the application system. Typical structure of a component includes the component interface and component implementation, and component interfaces and implementation are separated. As long as keeping the uniform interface standard in the application program, you can easily add or replace components in the system. Such as algorithm module in SmartMiner data mining system designed by Liu Junqiang, it used the component object model COM technology to structure, and provides registration mechanism for components by algorithm description library, and any algorithm module conformed to COM standard can be easily added to the system. In the MSMiner system Shi Zhongzhi and others studied and developed, various data mining core algorithms are implemented in the form of dynamic link library DLL, and can run dynamic loading

in the running process of system. The system also provides special algorithm management module, manages all kinds of mining algorithms through mining algorithms library, and provides the registration mechanism of the algorithm through the form of metadata.

Knowledge assessment and knowledge base layer. Before presenting the mining results to the user, knowledge evaluation can effectively remove the redundant and useless mining results, which havs important meaning to improve the usability of the system. Knowledge evaluation metrics include validity, novelty, potential usefulness and ultimately intelligibility. Qi Yanxia introduced in detail four kinds of combined ways of knowledge evaluation and data mining process. The knowledge model mined by data mining system can be stored in the repository for reusing after knowledge evaluation. To facilitate the sharing of knowledge model among different data mining system, DMG group (the data mining group) put forward prediction model markup language PMML. PMML is an XML-based language, which provided a unified definition and description standard for the prediction model the data mining created, made data mining systems of different vendors to follow the standard be easily share prediction model, and improved the reusability of the model and system scalability. Wettschereck introduced PMML application in the model interchange. Above implementation technology of various components of the centralized data mining system has been made detailed introduction. Now a lot of business data mining software based on the centralized structure has appeared and been widely used. More influential commercial software mainly include Enterprise Miner of SAS, IBM's Intelligent Miner and Clementine of SPSS, etc. Enterprise Miner realizes the integration with SAS Warehouse Administrator and OLAP (SAS/OLAP Server), which can realize the end-to-end knowledge discovery from the data proposing, seizing to solution. Intelligent Miner for Data supported for mining a variety of data source, such as traditional file, database, data warehouse and data center, etc. Clementine used data mining process model CRISP - DM (cross industry standard process for data mining), which can let users execute and manage the whole data mining work easily and effectively. At the same time, the three software provided the support for PMML 2.1, realizing the sharing of mining model.

Distributed data mining system

With the development and maturity of network technology and distributed database technology, the distributed database has been widely used, the centralized storage and management of the original data also gradually transformed into a distributed storage and management. The change of data storage way also is inevitable to promote the change of the data mining technology and its system structure. Due to the security, privacy and confidentiality of the data and network bandwidth limitations in the practical application, the method that first data set dispersed and stored are integrated into a database and then to be mined is not feasible, so the distributed data mining become the most feasible solution of data mining in a distributed database. The distributed data mining includes the following steps:

1) Subdivide data to be mined into P subsets, and P is the number of available processors, and each data subset is sent to each processor; 2) each processor run data mining algorithm in its local data subset, and the processor can run different data mining algorithms; 3) Combine of local knowledge various data mining algorithms find into a global, consistent findings. There are four key technologies in the distributed data mining: data set, parallel data mining, knowledge absorption and distributed software engine. Distributed data mining research mainly includes two aspects of the distributed data mining algorithm and distributed data mining system structure research. The current has appeared a lot of distributed and parallel data mining algorithms, such as the parallel algorithm for mining association rules CD (count distribution), DD (data distribution), and product data management (PDM), etc. In terms of distributed data mining system structure, it has also appeared a lot of architecture based on different techniques. For example, Zhang Xueming has studied distributed architecture parallel based on CORBA technology and adopting multi-thread parallel data mining mechanism, and Chen Gang has studied distributed data mining system structure based on mobile Agent technology. Hou Jingjun and others proposed a distributed architecture based on Web Services, which can realize data mining and protocol of large capacity

data in the distributed heterogeneous environment, and design the link service used for optical fiber network, and the framework can be used in the distributed data mining for Gigabyte large amount of data.

Conclusion

The current commercial data mining software has further promoted the popularization and development of data mining technology, but in practice there are still many problems and points need to continue to be improved, and research direction and development trend in the current main includes the following aspects: 1) enhance the visualization and interactive. A data mining system with good visualization and interactive function enables users to visually see and understand the definition and the implementation process of data mining tasks, reduces the blindness of user knowledge mining and production of a large number of independent mode in mining process, and improves mining efficiency of the system and user satisfaction and credibility to the mining result.

2) improve extensibility. Because the user's application environment is constantly changing, the scalability is very important for a data mining system. The system should support mining of data with a variety of sources and the scalability of mining algorithm, allowing the user to add the new algorithm according to the need. 3) combine with specific industry applications. With the development of application environment, the general data mining systems more and more cannot meet the needs of users. If the user don't know the character of the mining algorithm, it is difficult to draw a good model. Thus the data mining system should combine with a specific application, providing a complete solution for the application fields.

3) support mobile environment. The combination of data mining and mobile computing is a new research field. So that the data mining system which can mine the data generated by mobile system, embedded system and the ubiquitous computing devices is a new trend of development in the future.

References

[1] HAN Jia-wei,KAMBER M. Data Mining Concepts and Technigues [M] . FAN Ming,MENG Xiao-feng,trnsI. Beijing:China Ma-chine Press,2010. 305-307.(in Chinese)

[2] ZHOU Bin,LIU Ya-ping,WU Ouan-yuan. The design and impIementations issues of a data mining systems for eIectronic commerce[J]. Computer Engineering,2012,26 (6) :18-20.(in Chinese)

[3] WANG Jia-cai,CHEN Oi,ZHAO Jie-yu,et aI. VISMiner:An interactive visuaI data mining prototyped system [J] . Computer Engi-neering,2003,29 (1) :17-19.(in Chinese)

[4] LIU Kan,ZHOU Xiao-zheng,ZHOU Dong-ru. VisuaI data mining based on paraIIeI coordinates [J] . Computer Engineering and Ap-pIications,2013,39 (5) : 193-196.(in Chinese)

[5] NETZ A,CHAUDHURI S,FAYYAD U,et aI. Integrating data mining with SOL databases:OLE DB for data mining [A] . Pro 17th IntConf on Data Engineering [C]. HeideIberg:IEEE,2001. 379-387.

[6] ZHAO Zhi-hong,LUO Bin,CHEN Shi-fu. A structure of data mining system based on data warehouse [J] . Computer AppIications andSoftware,2012,19 (4) :27-30.(in Chinese)

[7] OIAN Wei-ning,WEI Li,WANG Yan,et aI. A data mining system for very Iarge databases [J]. JournaI of Software, 2012, 13 (8) :1540-1545.(in Chinese)