Research on Key Technologies of Data Mining System Based on Web Service

Wang Yi

Shandong Women's University

123114821@qq.com

Keywords: Web service; data mining system; key technology; research

Abstract. As more and more data accumulated, companies are eager to dig out the knowledge behind a large amount of data to support decision-making. The existing data mining tools such as IBM's IntelligenceMiner and SAS EnterpriseMiner, although they provide a comparatively rich mining function, these tools cannot dig out distributed and highly heterogeneous data on Internet/Extranet, and it is not effectively connected with the operating system integration and target is not strong. If enterprises use these tools, the invest is heavy, some of the mining work is not used, and the algorithm library is very difficult to be upgraded. This paper presents a data mining system framework based on Web services. It can be well integrated with the original operating system, and it can excavate the data in the distributed database. Additionally, it has the advantages of cross platform, cross language, easy deployment and dynamic management algorithm library.

1. Introduction

In recent years, people's capability of using information technology to produce and collect data has greatly increased. Therefore, a new challenge is brought out: in the era known as the era of information explosion, to make data truly become the resources of a company, it is necessary to make full use of it to serve for the company's own business decision-making and strategy for the development. Therefore, data mining and knowledge discovery (DMKD) technology came into being, and showed its strong vitality. Data mining refers to the process of extracting information and knowledge that hidden in it, people do not know in advance, but potentially useful from abundant, incomplete noisy, fuzzy, and random data. Data mining is a cross discipline in general, which brings together the researchers in different fields, especially scholars and engineering and technical personnel in database, artificial intelligence, mathematical statistics, visualization, parallel computing and so on [1]. Data mining technology from the beginning is the application-oriented. It is not only simple retrieval call for a particular database, but also refers to do micro and macro statistics, analysis, synthesis and reasoning of the data, and it attempts to find a correlation between events and activities, and even makes use of existing data to predict activities in the future to solve practical problems. Although some achievements have been made in the research of data mining, there are still some limitations. It faces the problems that data mining process requires specialist to actively participate in, integration of data mining system with existing operating system, mining distributed and heterogeneous data sources on Internet/Extranet, custom of algorithm library, sharing and maintenance and so on.

2. Research on Key Technologies of Data Mining System

2.1 Structure of Data Mining System Based on Web Service

According to the general process of data mining, the data mining system must include the data pre-processing part, the mining algorithm part and the mining model. Fig. 1 is the structure diagram of the system.

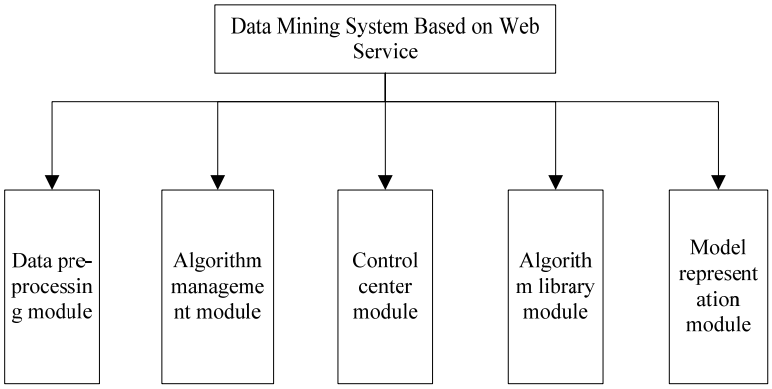

Figure 1 Structure Block Diagram of Data Mining System Based on Web Service

After the web services being introduced into data mining system, mining algorithms of data mining system are encapsulated into web services, and then form an algorithm library so as to achieve the dynamic management of the library. Therefore, data mining system construction based on Web services to construct must include data pre-processing module, algorithm management module, algorithm library module, model representation module and control center module controlling flow jump and coordinating various functional modules. In this way, it is able to establish a complete data mining system.

2.2 Data Pre-processing Module

The main function of the data pre-processing module is to be responsible for producing the mining data which can be used directly by the mining algorithm.

The data mining system, according to the data pre-processing method widely-used, in data pre-processing module, it adopts the data sampling and processing empty value, data segmentation, variable selection, and continuous data discretization and other data pre-processing function. The data flow chart of data pre-processing module is shown in Figure 2.

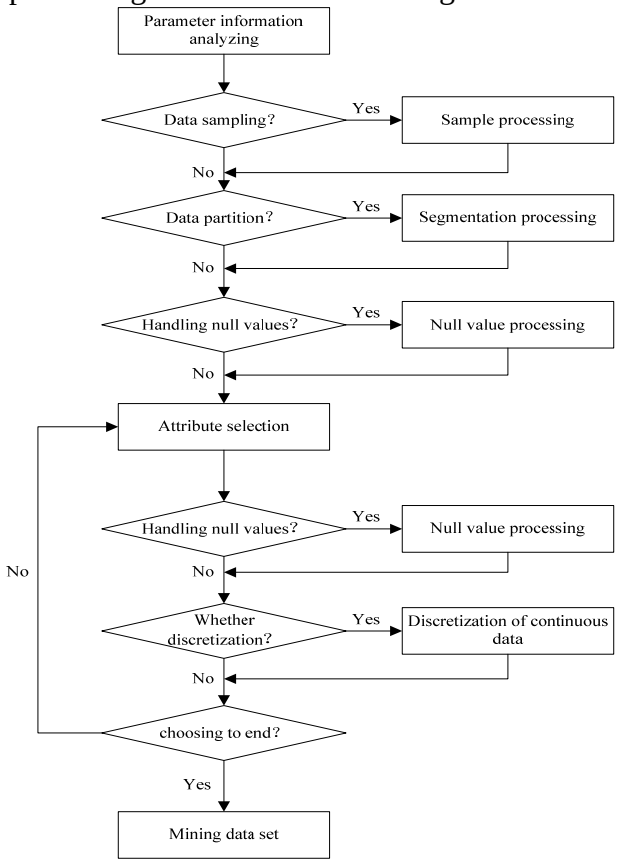

Figure 2 Flow chart of data pre-processing

(1) Data Sampling

When the amount of data used for data mining is very large, firstly the data set is sampled, and then proceed the data mining, which is a very effective way of data mining. At present, the main sampling methods are simple random sampling, equal distance sampling, sampling from the initial sequence, clustering sampling and stratified sampling.

The sampling module mainly completes the sampling operation of the input data set, and obtain a sample of the original data. For the large quantity data set, it is suggested to proceed data sampling. The data sampling module can perform simple random sampling, equal distance sampling, sampling from the initial sequence, clustering sampling and stratified sampling. For any kind of sampling method, it is possible to set the amount of sample record or the proportion of the amount of sample record accounted for the total amount of data, set seed value needed by producing random number, and classification variable used in cluster sampling and stratified sampling. The sampling module stores the sampling result to output data set to facilitate the use by the subsequent modules.

(2) Attribution Selection

Attribute selection refers to choose the attributes associated with the theme of the mining property from those of the data set, or discards the attributes containing a large number of null values data, decreases the input attributes number of data mining algorithm, reduces the dimension of the spatial data, which can greatly accelerate the speed of execution of the data mining algorithm and improve the quality of the data mining model.

(3) Continuous Data Discretization

Continuous data discretization refers to convert a class of continuous numerical data to discrete class data. Some data mining algorithms such as association rule algorithm and NavieBayes algorithm can only deal with the discrete data, so it is necessary to discrete the continuous data from the original data set. The commonly used discretization methods are the equal interval method and the equal frequency interval method. In addition, we can use the clustering analysis algorithm to disperse the numerical data.

(4) Null Data Processing

In the actual data set mining, part of the empty value records will appear in data sets. The existence of these null values can seriously affect the quality of the mining results model, so it is necessary to take effective measures to deal with the null data records.

There are two general ways to deal with null values: discard and fill in. When records where the null values exist have no effect on the produce of results of the model, users can choose to retain or discard null values records. When records where the null values exist have a great influence on the produce of the mining result, users must find a value to fill in the null value. Users can specify a value to fill in null value data, adopt the most frequent appeared values to fill in data, use the average value of property where the null value exists, or use the median value to fill in data (The null value data is numeric data).

2.3 Algorithm Management Module

(1) UDDI Registration Center

UDDI registration center consists of UDDI server. It accepts requests from the algorithm release sub module, WSDL documents of generation algorithm, and together with other information of mining algorithms (including the algorithm's service name, URL, etc.) are stored on the UDDI server. At the same time, it also accepted the request from the algorithm sub module, query the algorithm that has been published in the algorithm library and meet the conditions set by users, and return to the relevant information. The UDDI server mainly maintenance the basic information released to all Web services on it, playing the role o "Yellow page". UDDI registration center has two kinds, public registration center and private registration center. Algorithm providers can choose to publish the algorithm to the public registration center, but also build their own private registration center in the internal of enterprise.

(2) Algorithm Publishing Sub Module

Algorithm publishing sub module mainly realizes the algorithm functions that is encapsulated into Web service by algorithm providers, including receiving algorithm related information input by

algorithm provider, sends a release request to UDDI registration center, requests to add a new mining algorithm or delete, modify the mining algorithm has been submitted to algorithm library.

(3) Algorithm Finding Sub Module

Algorithm finding sub module is mainly to achieve function of the algorithm applicant to find the required mining algorithm in the algorithm library. According to the basic information that is provided by users and the algorithm needed to call, it sends a search request to the UDDI registration center to find the mining algorithm that can meet the conditions. If found, then return to the related information of the algorithm, including the name of the service, the information of the algorithm provider, Web services, URL and Web services, WSDL documents and other information. If not found, then return to the relevant information. Figure 3 is the structure of the algorithm management module [2].

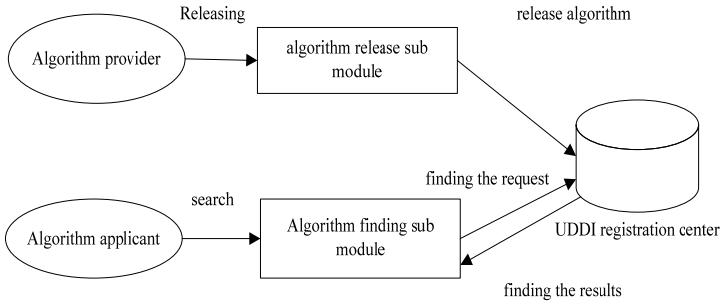

Figure 3 Structure of algorithm management module

2.4 Control Center Module

The control center module is the core of the whole data mining system based on Web services, which is responsible for the control of the whole system operation. Its main function is, according to the user's data mining requests, call the data pre-processing module mining data set to obtain mining data set, utilize the finding sub module in algorithm management module in the algorithm to call the relevant data mining algorithms, store the data mining model produced in the result database. In order to successfully complete the data mining, control center module will increase the corresponding data pre-processing requirements according to the special requirements of algorithm on data mining. For example, neural network algorithm can only deal with discrete data [3]. The mining algorithm has special requirement on data mining, and it is stored in a meta data repository. The control center module accepts and parses the request from the client, and calls other function modules of the data mining system to complete the data mining process together. Figure 4 is the data flow chart of the control center module.

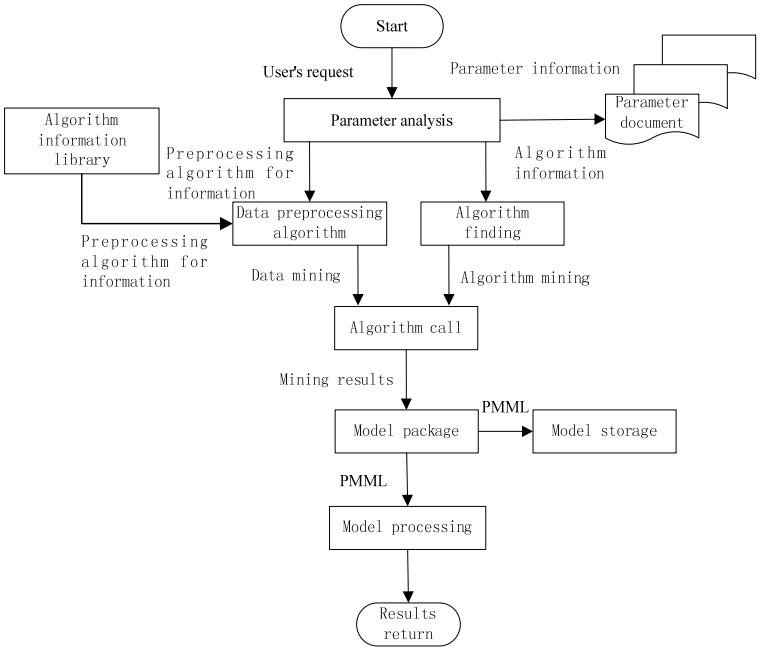

Figure 4 Flow chart of control center module

According to the request of the users, the control center module coordinates the work of each module in order to complete the whole process of data mining task.

When the calculation of the mining algorithm is finished, the control center module receives the SOAP message which contains the result and processes it into the mining model in PMML format. Store the mining model in PMML format along with the parameters file into the database [4]. At last, the data mining model of PMML document is processed into the corresponding representation form of the user to return to the user.

3. Realization of Data Mining System Prototype Based on Web Service

3.1 Development Tools

This prototype system mainly uses the Java language to realize. The development tool used is JBuilderX of Borland. BorlandJBuilder is the world's first cross platform Java integrated development environment, which can be used to build the Java application system matched to industrial standards and develop various application programs like Web, EJB, XML, WebService and database and so on.

3.2 Control Center Module

Control center module is the scheduling center of prototype system. Control center is the core of the data mining prototype system. It is the key of the whole data mining system to coordinate the data communication of each data mining model and call the different function modules to complete the data mining.

3.3 Advantages of System Framework

The data mining system adopts the framework oriented to service, uses the technology of Web services, so enterprises can easily customize their mining algorithm library according to their business needs, and realize dynamic management and upgrade of mining algorithm library. Compared with previous data mining algorithm in the system and other functional module closely coupled structure, the data mining system has higher superiority: algorithm library dynamic management, platform independent and language independent, algorithm library distributed deployment, and on-demand service mining [5].

4. Conclusion

After fully research on the technology of Web services, the service-oriented architecture is introduced to construct data mining system, and according to the characteristics of web service, do further research on the construction of the key technologies of the data mining system, and finally realize the prototype based on Web service data mining system. Compared with the previous data mining tools, the data mining system has the advantages of easy for dynamic management of data mining algorithm library, which greatly reduces the cost of enterprise development and deployment of data mining system [6]. Due to limited by development time, level and environmental, mining algorithm kinds and quantities in algorithm library must be further enriched; visualization of model results mining is yet to be further studied. It is believed that in the near future data mining technology will be mature and widely used as the same as the database technology.

References

[1] Lin, Tsau Young, Yiyu Y. Yao, and Lotfi A. Zadeh, eds. Data mining, rough sets and granular computing[J]. Vol. 95. Physica, 2013.

[2] Wu X, Zhu X, Wu G Q, et al. Data mining with big data[J]. IEEE transactions on knowledge and data engineering, 2014, 26(1): 97-107.

[3] Tang Q Y, Zhang C X. Data Processing System (DPS) software with experimental design, statistical analysis and data mining developed for use in entomological research[J]. Insect Science, 2013, 20(2): 254-260.

[4] Zhang G L, Riemer A B, Keskin D B, et al. HPVdb: a data mining system for knowledge discovery in human papillomavirus with applications in T cell immunology and vaccinology[J]. Database, 2014.
[5] Sakaeda T, Tamon A, Kadoyama K, et al. Data mining of the public version of the FDA Adverse Event Reporting System[J]. Int J Med Sci, 2013, 10(7): 796-803.
[6] Larose D T. Discovering knowledge in data: an introduction to data mining[M]. John Wiley & Sons, 2014.