Cloud Computing Based Data Mining of Medical Information

Lihua Wang[1, a *], Ze Zhang[2,b]

[1]College of Electronic Information and Engineering, Inner Mongolia University, China

[1]Affiliated Hospital of Inner Mongolia Medical University, China

[2]College of Electronic Information and Engineering, Inner Mongolia University, China

[a]ngdwlh@163.com, [b]zhangzeimu@163.com

Keywords: Cloud computing; Big data; Data mining; Medical information.

Abstract: Challenges of data mining in the medical industry in an era of big data are introduced herein. Combined with cloud computing, a frame of medical cloud data mining platform built in the ecological environment is proposed, with functions at various levels being introduced in detail, including the base level, the platform level, the function level and the service level. The research aims to provide useful information for big data analyses and mining in the medical industry.

Introduction

With the development and extended application of big data in the medical and life science research, their number is incredibly large with an extensive range. For instance, a CT medical image contains data of 100MB approximately, the size of a genome sequence file is about 700MB, and a standard pathological image contains nearly 4G data. Multiplying these medical data by population and average life expectancy, only a community hospital or a medium-sized pharmaceutical companies may generate and accumulate several PB structured and unstructured data. Consequently, it is of great significance to develop an efficient platform for the data mining of the medical information.

This research aims to propose a cloud computing based platform for the data mining of the medical information.

Definition of Cloud Computing

Cloud computing is a kind of model which is a product of scientific and technical information background. Cloud computing possesses real-time and convenient access to the share resources, which reduces the workload of the resource management and interaction with the service provider[1]. Cloud computing has the following characteristics:

(1) Data providers are responsible for the investment and maintenance of the infrastructure construction;
(2) Service provided by the infrastructure provider may be shared by multiple users;
(3) The users may receive appropriate data service according to their needs;
(4) Service space is scalable and flexible, and users can may get unlimited service by following certain requirements;
(5) Service quality is guaranteed by the provider.

Data Mining of Medical Information

Definition of Data Mining

Data mining is to discover some of the phenomena as well as the rules and knowledge through data calculation, which has been paid more and more attention in the field of computer science. Data mining combines the data analyses and data classification technologies, with a large number of random data being investigated. It may effectively extract implicit information and knowledge, and provide reference for the real decision-making[2].

Data mining is mainly divided into two types: one is description type and the other is prediction type. The descriptive data mining is to obtain general rules and features from the database through data calculation, and to provide data support. The predictive data mining is supported by a database, and provides predictions for realistic phenomenon through data calculation and resultant general rules[3]. Data mining has the following features:

(1) Class description. Data mining can take advantage of the correlation of the data or their classification to characteristically process and distinguish the data. Data characterization may undertake the data query through the database, and finally summarize the data features.
(2) Prediction and modeling. Prediction can be divided into two types: one is classification which mainly uses discrete target variable, and the other is the regression which is applicable to continuous traget variables. Both are able to independently predict the target and to derive a model.
(3) Correlation analysis. It is mainly used for describing strong correlation characteristics of the data. Through the data correlation analysis, it is possbile to clarify conditions of the high frequency occurrence of attributes in the data set area, and to reveal the relationship between the elements of the data set.
(4) Anomaly detection. Abnormal data can be detected based on data analyses. The anomaly detection algorithm has a lower error rate, which is used primarily for network security.

Data Mining Algorithms

Data mining is widely used in the medical field and has many algorithms, such as clustering algorithm, decision tree algorithm, association rule algorithm and Bayesian analysis, etc. The decision tree algorithm and Bayesian network are investigated in detail below.

Decision tree algorithm is used to classify the data in the given rules, and to find the rules and process of the classification of the dataset[4]. The essential work of the decision tree algorithm is to construct the decision tree of small-scale and high-precision, which is divided into three stages: firstly, the data set is divided into training set and test set; secondly, a decision tree is generated by the trained rules; thirdly, the rules generated in the process are checked through calculation of the decision tree, with the degree of data errors being obtained [5]. The decision tree possesses many advantages in terms of more intuitive and easily-understood data, convenient preparation process, being more suitable for processing text data with small volumn of computing and high speed, as well as usage in big data analysis.

Bayesian network is one kind of probabilistic reasoning based on Bayesian computing network, with the aim of solving uncertainty and imperfection issues of the data. Bayesian network has a strong ability to adapt the data noise. It can be divided into the following steps for developing a Bayesian network: firstly, to determine the data variables involved in the calculation; secondly, to establish a network structure based on dependencies between the variables; thirdly, to determine the probability distribution among the variables through data calculations; fourthly, to optimize the Bayesian network in accordance with the data generated by the calculation. Bayesian network has advantages in terms of processing noise data, providing users with a complete associated probability model and interpretation of training results on the data[6].

Data Mining Process

Data mining process has cycle states, i.e. a data mining model can be an iterative process. Data mining process can be divided into the following steps:

(1) Problem determination. The purpose of data mining is to discover potential information that can accurately reflect the phenomenon through analysis and calculation of massive data. Therefore, clarifying the information that needs to be mined is the first thing for the data mining[7]. In determining the own problem, on one hand practical needs should be clear, on the other hand a more appropriate algorithm should be selected by comparison and screening of various algorithms.

(2) Data preprocessing. Data preprocessing mainly contains the supplement of data, elimination of data duplication and data type conversions and so on.
(3) Data flow. Data flow technology includes calculations of the minimum and maximum values, the standard deviation and the mean deviation. It may contribute to the structural stability and accuracy of the information to judge.
(4) Model generation. The last step of data mining is to generate one or more mining models. According to the given mining problem and clear data mining object, potential rules within the data are mined by selecting appropriate algorithms.

Application of Big Data Mining in Medical Industry

Clinical Decision Support System

Big data analysis techniques will make clinical decision support systems more intelligent thanks to increasingly strengthened capabilities for unstructured data analyses. For instance, image analysis and recognition technology may be utilized to identify medical imaging data, or to provide doctors with treatment recommendations through medical expert database developed by data mining of medical literature[8]. In addition, clinical decision support systems can also transfer most of the work in the medical process to nurses and assistant doctors, so that doctors can be freed from simple consultation work that takes too much time, and the efficiency of diagnosis and treatment can be significantly improved.

Medical Image Data Mining

Cloud computing based distributed data mining platform frame is a high-scalability and high-performance parallel computing programming model.Medical images are formed by means of imaging equipment and through various characteristics of transmission, reflection, and absorption of different human organs and tissues against the radiation, ultrasound, light scattering, etc.With the development, medical imaging has provided an important basis for the diagnosis of disease, preoperative decision-making and postoperative review and so on. But, current medical imaging diagnosis is still largely dependent on the clinical experience and subjective judgment of the doctors, so it has the practical significance to assist doctors in clinical diagnosis to make use of the clinical imaging data and expert's clinical diagnosis experience and knowledge, using computer science and technology quickly and detecting lesion area in medical image accurately. Medical image data mining flow chat is shown in fig1.

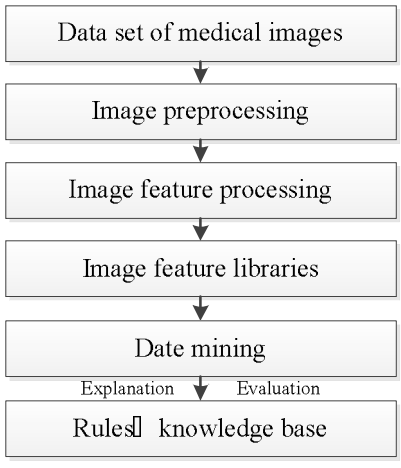

Fig. 1 General process of medical image data mining

Public Health

Big data mining can improve public health monitoring level. On the basis of national electronic patient records database, public health departments may rapidly detect infectious diseases, conduct comprehensive surveillance, and quickly respond through integrated disease surveillance and response procedures. This will bring many benefits, including reduced medical claims expenses, reduced infection rates of infectious diseases, and more quick detection of new infectious diseases by health departments. By providing accurate and timely public health advice, it may lead to a substantial increase in public awareness of health risks, and reduce the risk of infection diseases[9].

Cloud Computing Based Medical Information Platform Hadoop

Platform Hadoop

With the continuous expansion of the Internet data size, the data processing requirements continue to increase, and traditional network framework cannot afford such a large number of data resources. Therefore, the emergence of new network technology framework makes big data processing and storage in the medical information possible. Hadoop is one of mainstream cloud platform solutions, which can internally form data cluster through computer, use the computing resources in the cluster for data processing, and form a distributed framework. Hadoop system possesses strong scalability, and is highly fault-tolerant, with a programming model that can help users easily write program code. Different from traditional SQL databases, Hadoop extended upwards through functional programming. And, Hadoop uses offline processing mode rather than online one to provide users with high-level services. Hadoop system frame diagram of the various components is shown in Fig. 2.

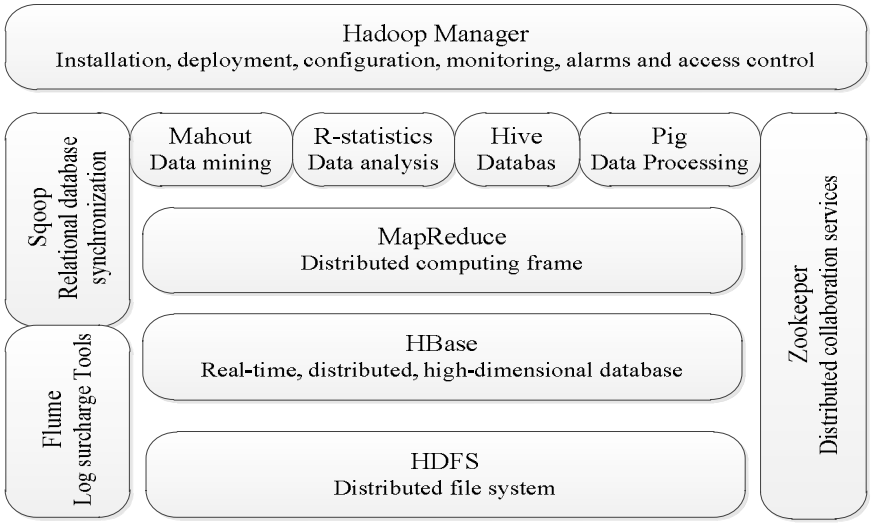

Fig. 2 Components and system frame diagram of Hadoop

Characteristics of Platform Hadoop

Hadoop platform is an open source framework, developed by using the Java language, and therefore possesses strong portability. Hadoop platform has the following characteristics:

(1) Scalability. Platform Hadoop can be easily extended, without making any changes to the structure.

(2) Low cost. The Cluster of Platform Hadoop is composed by computers, with more nodes and lower cost.

(3) High efficiency. The system consists of a distributed framework and a parallel MapReduce framework, and may process the data more quickly.

(4) Security. Platform Hadoop has strong reliability, with MapReduce task monitoring technology being used; Platform Hadoop platform is capable of reconnection if task fails.

Summary

Cloud computing is a new way of shared infrastructure. Cloud computing based medical information service platform will be achieved mainly through improvement and evolvement of existing frameworks. Core technologies of cloud computing such as virtualization and replication technologies enable low cost high demand service level agreement possible, and medical information services will have new development directions thanks to cloud computing. Cloud computing has become the driving force behind the development of the big data industry, and successful application cases can be seen in all kinds of business. The medical big data platform developed herein based on the ecological environment may provide new information and idea for big data analysis and mining in the medical industry.

References

[1] W. Fan, D. Zhao, S. Wang, A cloud computing-based implementation of regional medical information sharing, J. Military Medical Sciences. 4 (2015) 257-260. (in Chinese)

[2] Y. Song, Y. Wang, W. Cao, Investigation on medical big data mining based on cloud computing, J. Military Medical Sciences. 2 (2015) 11-13. (in Chinese)

[3] Y. Tang, J. Liu, H. Gan, Design and study on regional medical and health information system (RHIN) by using cloud architecture, J. Chinese Journal of Health Informatics and Management. 10 (2013) 96-104. (in Chinese)

[4] J Zhu, Investigation on platform frame and key technologies of data mining based on cloud computing, J. Computer CD Software and Applications. 21 (2014) 111-113. (in Chinese)

[5] H. Gao, L. Xiao, D. Xu, Z. Sang, Medical data mining platform based on cloud computing, J. Journal of Medical Informatics. 5 (2013) 7-12. (in Chinese)

[6] D. Wang, Research on Optimization of Apriori Algorithm on Cloud Computing and Medical Data, D. Beijing University of Posts and Telecommunications. 2015. (in Chinese)

[7] M. Xu, Research and Construction of Hospital Information Technology Based on Cloud Computing, D. Xiamen University. 2014. (in Chinese)

[8] J. Ji, Design and Achievement of a Data Mining Platform Frame based on Cloud Computing, D. Qingdao University. 2009. (in Chinese)

[9] Q. Li, Research on Design and Realization of Health Information Center basing on Cloud service, D. South China University of Technology. 2011. (in Chinese)