RESEARCH ARTICLE OPEN ACCESS

Application of data mining with on cloud computing

Dr.S Krishna Mohan Rao1, Jagannath Ray2, Mohini Prasad Mishra3
[1,2]*Associate Professor, Department of Computer Science Engineering, Gandhi Institute For Technology (GIFT), Bhubaneswar*
[3] *Assistant Professor, Department of Computer Science Engineering, Gandhi Engineering College, Bhubaneswar*

ABSTRACT. With the rapid increase of data storage capacity, massive data processing and massive data calculation has become an important problem in the field of data mining. Cloud computing is good at dealing with large-scale data and large-scale computing. If the data mining algorithm can be cooperated to the cloud computing platform, the large computational problems in the field of data mining will be solved. In this paper, the author introduces the basic characteristics and process of cloud computing and data mining, and summarizes the various methods of data mining in detail, including neural network method, genetic algorithm, decision tree method, statistical analysis, rough set method. Finally, the application of data mining based on cloud computing is summarized.

I. INTRODUCTION

The data processed by modern society is massive. Before the advent of cloud computing, when conducting data mining in the past, the high-performance machine or a larger computing device were expected to deal with it. In the context of massive data, the data mining process requires a good development environment and application platform. In such circumstances, the use of cloud-computing-based approach for data mining is more appropriate. And because of the lack of the current parallel classification algorithm and large-scale data sets being increasingly large, the traditional data mining system cannot be used for efficient mining and utilization. Therefore, how to improve the parallelism and efficiency of the algorithm is the urgent issue to be solved. By sorting out the development and application of data mining based on cloud computing, we hope to provide some help for future data mining research.

II. THE CONCEPT OF CLOUD COMPUTING

Cloud Computing [1] is an Internet-based computing approach, in which shared hardware and software resources and information can be provided to the computer and other equipment on demand. Cloud computing describes an Internet-based new IT service addition, usage, and delivery model that provides dynamic, scalable, and often virtualized resources over the Internet. Cloud Computing [2,3] is a computing platform which is distributed in large-scale data centers, and can dynamically provide various server resources to meet the needs of research, e-commerce and other fields. Cloud computing is the development of Distributed Computing [4], Parallel Computing [5], and Grid Computing [6], and the integrated evolution result of Virtualization [7], Utility Computing [8], IaaS (infrastructure as a service), PaaS (platform as a service), and SaaS (Software as a service).

III. THE FEATURES OF CLOUD COMPUTING

3.1. Verification of the Finite Element Model

Cloud computing provides the most reliable and secure data storage center, users need not worry about data loss, virus invasion and other troubles. The “Cloud” uses data replica fault tolerant and interchangeable compute nodes to protect the high reliability of service so that using cloud computing is more reliable than using local computers.

3.2. Virtualization Technology

The most important feature of the existing cloud platform is virtualization technology, which can realize virtualization management, dispatch and application of hardware resources of servers through virtualization technology. Cloud computing allows users to access applications at any place and from a variety of terminals. The requested resource comes from the cloud, not the fixed tangible entity. Applications are running somewhere in the cloud, but in fact, users don't need to understand nor worry about where the application will run. With just one laptop or a mobile phone, you can do everything you need, including even supercomputing, through web services.

3.3. Flexibility of Service

Users can customize according to their own needs in the cloud platform personalized services,
applications and resources, cloud platform is based on users' demand to deploy resources, services, the corresponding calculation capability and application. Users can also manage customized services, unsubscribe, deleting certain services, and so on.

3.4. Versatility

Cloud computing is not specific to certain applications, and can be built with the support of the "cloud". The same "cloud" can support different applications at the same time.

IV. THE CONCEPT OF DATA MINING

Data mining [9] (DM) is the process of extracting the implied, unknown in advance but potentially useful information and knowledge from a large number of incomplete, noisy, fuzzy, random data. With the rapid development of information technology, the amount of data accumulated by people rapidly increase, often is calculated by TB, how to extract useful knowledge from the massive data has become a problem, which must be solved. Data mining came into being for adapting to this need and became a data processing technology. The basic process of data mining is shown.

V. 5. THE METHOD OF DATA MINING

5.1. Neural network method

Because of its good robustness, self-organization adaptability, parallel processing, distributed storage and high fault tolerance, neural network is very suitable for solving the problem of data mining, that it has been paid more and more attention in recent years. The typical neural network model is divided into three categories: feedforward neural network model for classification, prediction and pattern recognition, represented by perceptron, Back propagation (BP) and function network. The feedback neural network model taking the hopfield discrete model and continuous model as the representative for associative memory and optimization calculation, respectively; A self - organizing mapping method for clustering, represented by the art model and the koholon model.

5.2. Genetic algorithm

Genetic algorithm is a kind of stochastic search algorithm based on biological natural selection and genetic mechanism, which is a bionic global optimization method. The genetic algorithm has the implicit parallelism, which is easy to combine with other models and other properties so that it is applied in data mining.

The application of genetic algorithm is also reflected in the combination of neural networks, rough set and other technologies. Such as the use of genetic algorithm to optimize the neural network structure, under the premise of not increasing the error rate, delete the redundant connections and hidden layer unit; genetic algorithm and bp algorithm are combined to train neural network, and then extract the law from the network.

5.3. Decision tree method

Decision tree algorithm is one of the most basic and most commonly used algorithms in data mining, which is similar to the tree structure of a flow chart. Each internal node in the tree epresents a test of a property, and each branch represents a test result, while each leaf node represents a classification. The basic algorithm of decision tree induction is the greedy algorithm, which constructs the decision tree in the way of top-down recursion [10]. The construction of decision tree algorithm usually goes through two stages: tree construction stage and tree pruning stage.

5.4. Statistical analysis

There are two relationships between database field items: function relationships (deterministic relationships that can be represented by function formulas) and correlations (which can not be represented by functional formulas, but still relevant deterministic). The analysis of them can be done by statistical method, that is, using the principle of statistics to analyze the information in the database. Common statistics, regression analysis, correlation analysis and difference analysis can be used.

5.5. Rough set method

In 1980s, Pawlak proposed the concept of Rough Set [11,12]. Rough set is a mathematical tool with fuzzy and uncertainty, it is commonly used to study data expression, study and induction, and is widely used in many fields such as data mining, knowledge discovery, uncertainty reasoning, granular computing. The application of rough set theory to classification algorithm can help to find inaccurate or data structure relationships existing in noisy data. However, rough sets only can process discrete data. If it is a continuous data object, it needs to be discretized first and then processed by rough sets. Rough sets can be used to reduce attributes, correlation analysis and other operations, according to the upper approximation and lower approximation to determine the data set. The lower approximation set contains data samples that belong to the data set, and the upper approximation set contains data samples that definitely do not belong to the data set.

VI. APPLICATION OF DATA MINING BASED ON CLOUD COMPUTING

Data mining technology is the application service from its emergence, and it has a wide range of applications in various fields. The application is especially broad in the financial, retail, tele communications, electronic engineering, aviation, medicine, transportation and other fields. The typical application of data mining in the field of commercial includes database marketing, customer classification, background analysis, market behavior analysis and customer churn analysis, credit rating, fraud and so on. Specific application areas are shown.

VII. CONCLUSION

Because the cloud computing with its distributed computing platform provides a powerful computing, the combination of cloud computing and data mining has a huge advantage and potential. The application of cloud computing to data mining can provide solutions for more and more massive data mining, which has become the trend of data mining industry development. This paper summarizes the basic concept and principle of cloud computing and data mining in detail, as well as the basic methods of data mining. Finally it sorts out the related application of cloud based data mining. We hope that this paper will provide a new direction for future data mining research

REFERENCES

[1]. R Buyya, CS Yeo, S Venugopal. Market-oriented cloud computing: vision, hype, and reality fordelivering IT services as computing utilities [C]. Proceedings of the 2008 10th IEEE International Conference on High Performance Computing and Communications, 2008,5-13.

[2]. A Weiss. Computing in the clouds. networker, 11(4): 16-25, Dec. 2007.

[3]. R Buyya, CS Yeo, S Venugopal. Market-oriented cloud computing: vision, hype, and reality fordelivering IT services as computing utilities [C]. Proceedings of the 2008 l0th IEEE International Conference on High Performance Computing and Communications, 2008,5-13.

[4]. H Attiya, J Welch. Distributed Computing: Fundamentals, Simulations and Advanced Topics [M].John Wiley & Sons, 2004, 4(2): 197-224.

[5]. V Kumar. Introduction to parallel computing [M]. Addison-Wesley Longman Publishing Co. Boston, MA. 2002

[6]. F Berman, G Fox, AJG Hey. Grid computing: making the global infrastructure a reality [M]. John Wiley and Sons. 2003

[7]. P Barham, B Dragovic, K Fraser, S Hand, T Harris. Xen and the art of virtualization [C]. ACM SIGOPS Operating Systems Review, 2003, 37(5): 164-177.

[8]. CS Yeo， MD de Assuncao, J Yu, A Sulistio. Tility Computing and Global Grids. Grid Computing and Distributed Systems Laboratory, The University of Melbourne, Australia, April 13,2006.

[9]. J Han, M Kamber. Data mining: concepts and techniques [M]. Morgan Kaufmann Publishers, 2001.

[10]. JC Platt. Fast training of support vector machines using sequential mining optimization [R] Scholkopf B, Buuges CJC, Smola AJ, editors, Advances in Kernel Methods-Support Vector Leraning. MIT Press, 1999, 185-208.

[11]. Y Wu, K She, W Zhu, et al. A Web text filter based on Rough set weighted Bayesian [C]. 8thIEEE International Symposium on Dependable, Autonomic and Secure Computing, 2009:241-245.

[12]. I Naohiro, Y Takahiro,YG Bao. Rough set based Learning for classification [C]. 20th IEEE International Conference on Tools with Artificial Intelligence, 2008: 97-104.