Study and Application of Big Data Mining Based on Cloud Computing

Jinwei Luo[1], Chunfei Li[2] and Fuping Huang[3]

[1]Guangdong innovation and technical College, Dongguan Guangdong, 523960, China

[2]Guangzhou Sontan Polytechnic College, Guangzhou Guangdong, 511370, China

[3]Guangzhou Zhujiang College of Vocational Technology, Guangzhou Guangdong, 511370, China

Keywords: Cloud computing, Big data mining, Mode.

Abstract. Considering the constant development of modern information technology, various kinds of data possessed by people have increased in a geometric and explosive way.Big data mining technology is a key technology for obtaining relevant information and has become a focus of study in information bound. As data to be mined by us increase at an exponential speed, centralized serial data mining technology in the traditional sense is no longer applicable to the reality. Therefore, being committed to improving the ability of data processing of big data mining algorithm and improving the efficiency of data processing has become a vital research topic. This paper describes relevant contents of cloud computing and big data mining, the necessity of applying cloud computing in big data mining and important big data mining technologies under cloud computing and analyzes big data mining model from the perspective of cloud computing.

Introduction

With the entry of China into internet + era, network technology especially modern mobile internet technology has developed rapidly. The volume of various kinds of data are increasing at a high speed at every moment. Facing the information with amazing quantity, finding the content that users are interested in becomes a difficult thing. Therefore, big data mining wins great popular in all walks of life. Meanwhile, cloud computing technology becomes a popular field of technical research and development, which can make big data mining system produce a brand new development trend, form a good data mining system and then make it possible to deal with massive data in reality. Cloud computing can be considered as the product of mutual penetration between computer technology in the traditional sense and modern new network technology. Cloud computing is not a simple compute mode in the traditional sense. Instead, it is a brand new compute mode and the complex of such forms as data storage, backup and network resource allocation. Previous big data mining platform is a compute mode established based on database and using data information that has been collected so as to find out relevant information hidden in differential data. Big data mining in the traditional sense must be established based on massive data and conduct a lot of data access and calculation work. During big data mining, it will produce a lot of consumption and occupy a lot of computing and storage resources. Data with rapidly increasing size cause the failure of previous big data mining technology to deal with them. In this situation, the author uses cloud computing mode to put forward a better countermeasure for more data mining.

Elaboration of relevant contents of cloud computing and big data mining

Cloud computing is one of the commercial compute modes under internet. In terms of the concept of cloud computing, there is no consistent statement in the academic world. The one accepted by most people is the concept defined by the National Institute of Standards and Technology, i.e. cloud computing is one of the modes with payment based on usage amount. This mode provides network access available and convenient and the shared pool with the ability to allocate corresponding computing resources can be entered. Resources herein mainly cover network, server, application software and service etc. Such resources can be provided quickly. It is only necessary to input very limited management force. Cloud computing has a large-scale shared pool of computing resource

type that can be allocated. This means that cloud computing is resource allocation technology or service conducted for the shared pool. It is characterized by the ability to provide computing resources above very quickly so as to reduce the work amount of customers engaged in management activities. Cloud computing assigns computing tasks to the shared resource pool composed of many computers or servers, thus greatly improving the utilization rate of resources, comprehensively improving abilities such as computation and storage and possessing ideal expansion. Virtualization feature of cloud computing makes users unrestricted by their geographical position and even terminal equipment. This means that various application services requested can be obtained as long as users can surf the internet. In other words, users can benefit and obtain the required application service as long as they have any computer terminal equipment that can surf the internet. Cloud computing has universality. Cloud computing platform can produce a lot of applications. Therefore, users are unrestricted in application and can use multiple different applications under cloud platform. Cloud computing has super-large scale and high expansion. Scale expansion of cloud will not influence actual quality of user service. Currently, cloud computing platform has a large scale. For example, cloud computing of USA Google Company has millions of professional servers. Clouding computing is characterized by high reliability and profitability. Relevant technologies such as multi-replica fault tolerance guarantee high reliability of services. Cloud computing mainly uses cheap nodes to constitute cloud and integrate automation and centralized management system. Compared to data management cost of enterprises in the traditional sense, it has good economical efficiency.

Big data mining is mainly one of the important steps in database knowledge discovery. Big data mining is the process of mining massive data mastered currently with a specific algorithm and finding out practical information or knowledge. Big data mining generally aims at solving various demand problems and meanwhile handling prominent problems such as soft data management involved properly. For information bound, big data mining is an important link forming value. Only by transforming data into information or knowledge with practical value can it have real business development value. Big data mining in the traditional sense is established based on giant database. Database system is required to provide good support such as storage, index and query. Good computing technology is an important support for massive data processing. Therefore, it has significant influence on processing efficiency. Considering the continuous expansion of network scale and the coming of mobile internet era, the scale of various data is increasing at a fast speed and there are more and more requirements for big data mining. This causes some problems of previous big data mining technology. First, the efficiency of big data mining is low. Previous big data mining technology based on database cannot complete various computing tasks very well in front of current massive data. Second, in the situation of massive data scale, previous big data mining technology must have higher software and hardware for support. Third, big data mining technology based on data system can hardly continue to provide more support for the improvement of the ability of mining algorithm. Once an algorithm is restricted by system architecture, further development of big data mining will be influenced.

Necessity of using cloud computing in big data mining

As data volume has increased rapidly in a geometric way in recent years, the low value density of data becomes more and more prominent. In the current era of big data, data become more and more important. However, data mining technology must be used so that great value of big data can be reflected in massive low-value density data. Big data mining often needs to use a lot of data so as to get valuable information and improve the model. Long operation time is often consumed to implement multi-frequency data access to massive data. The conflict between the complexity of big data and relatively limited computing power of system becomes more prominent. Previous single-computer system has problems such as low speed, low efficiency and high energy consumption. However, cloud computing has such features as the ability to implement dynamic resource allocation, virtualization and high availability and therefore can meet the requirement of mining calculation very well. The formation and development of big data mining technology cannot exist separately without cloud computing technology. Cloud computing distributes originally

complicated computing tasks to the cloud constituted by computer system and allocates abilities such as calculation, storage and application service to users according to demand so as to improve the acquisition rate of data. Big data mining extracts valuable information from massive, incomplete and random data after screening and optimization through processing. As these data have a complicated situation, large storage and calculated amount should be used. Big data mining platform based on cloud computing technology can solve the problem well. It can actually control operation and storage costs, further improve the efficiency of data mining and actually break various restrictions in previous big data mining.

Important big data mining technologies under cloud computing

Generally speaking, big data mining technologies under cloud computing mainly cover distributed parallel technology and data mining algorithm which will be described respectively below. First, distributed parallel technology. The author thinks that the most important function of cloud computing is storage and parallel calculation of various distributed files. A prominent function of the former is to improve the processing rate of data and then meet various requirements of parallel calculation. The earliest distributed file system was GFS system developed by USA Google Company. Subsequent systems such as HDFS and KFS were developed based on GFS system. Currently, systems above have been widely used in commercial and academic fields. In the field of parallel calculation, MapReduce programming method of USA Google Company is most widely used. Data are packaged after encoding processing of various problems such as data distribution, task implementation, data tolerance and bandwidth delay. Users only need to invoke and execute them. However, this method is not application in the field of cross correlation coefficient data statistics and comprehensive development system is not established. Second, data mining algorithm. The algorithm mainly covers such research fields as statistics, artificial intelligence and modeling. It is a principal technology in big data mining. General methods include statistical analysis, decision-making tree and neural network etc.

Analysis on big data mining model from the perspective of cloud computing

Big data mining technology based on cloud computing uses the storage capacity and distributed parallel capacity of cloud computing. It can be considered as big data mining model based on cloud computing. There are mainly three layers of structure and five modules. Three layers of structure refer to top layer, interlayer and data center layer, which will be analyzed respectively below. The first layer is top layer. Top layer mainly involves workflow and user interface subsystem. It is generally user-oriented. The main function of the former is to help users to form various tasks related to data mining. The main function of the latter is to promote the realization of user interaction function. Users can design reasonable parameters in cloud computing input modules, use the most suitable data mining algorithm for processing, then use MapReduce platform to implement big data mining work and finally show the result to users in a visualized form. The second layer is interlayer. It is the central part in data mining system under cloud computing. It mainly includes two subsystems - data pretreatment and parallel data mining. Considering that MapReduce model under cloud computing environment focuses on data with the same type and structure, data pretreatment subsystem needs to process various irregular big data in advance. Its result is the input of data mining algorithm. General data pretreatment modes cover parallel data cleaning, transformation, extraction and loading. After pretreatment of big data, the proportion of noise data and useless data will reduce on a large scale. Therefore, the efficiency of data mining is greatly improved. The latter is one of the most important modules in computer big data mining system. Currently, a lot of excellent data mining algorithms have emerged continuously. However, as MapReduce is an algorithm model in cloud computing platform, it is difficult to directly use algorithms above in cloud computing platform. This requires integration and transformation of various existing algorithms so as to promote the realization of cloud computing optimization. The third layer is data center layer. This layer is in the lowest position of the system. Its main function is to store massive big data and implement distributed parallel processing of

data on this basis. This layer must retain multiple copies at the same time so as to guarantee the security and high availability of big data. As cloud computing implements parallel mode, it can give a response rapidly even when many users make relevant requests at the same time. Therefore, big data mining application under cloud computing can greatly improve the efficiency of data processing. Future research focus is to reduce the scan times of database and then improve the quality of big data mining.

Conclusion

In conclusion, the research, development and application of big data mining system under cloud computing platform have obtained a good efficiency. The system has its unique advantages such as high efficiency and large amount of data processing. However, as cloud computing is still in the initial stage of development, there are many problems to be solved. It is required to not only overcome problems in safety performance of cloud computing software comprehensively, but also endeavor to overcome the possible problem of uncertainty of algorithm and result in the process of big data mining. Therefore, it is required to pay attention to combine with the reality, improve the degree of personalization and universality of design, constantly improve the degree of encryption of private data and then achieve better efficiency of big data mining in the process of positively establishing big data mining system under the premise of cloud computing.

Acknowledgments

Topic name of this paper: Study on Teaching Resource Construction of Vocational Education Based on Cloud Platform, No. GDJY-2014-B-b248.

References

[1] Wang Peng. Cloud Computing and Big Data Technology. Posts and Telecom Press, 2014.

[2] Cheng Lin. Study on System Architecture of Data Mining Based on Cloud Computing, *Electronic World*, 2012(21).

[3] Ding Yan, Yang Qingping, Qian Yuming. Study on Framework of Data Mining Platform Based on Cloud Computing and Its Key Technologies, *ZTE Communications*, 2013(1).

[4] He Qing, Zhuang Fuzhen. Big Data Mining Platform Based on Cloud Computing, *ZTE Communications*, 2013 (4).

[5] Gao Hansong, Xiao Ling, Xu Dewei, Sang Ziqin. Medical Big Data Mining Platform Based on Cloud Computing, *Journal of Medical Informatics*, 2013 (5).

[6] Zheng Miaoshi. Study on Architecture of Data Mining Platform Based on Cloud Computing and Its Key Technologies, *Information Communication*, 2014 (8).