Study and Application of Big Data Mining Based on Cloud Computing

Jie Shao

Hubei Engineering Vocational College, Huangshi Hubei, 435000, China

Keywords: Cloud computing, Big data mining, Research and application

Abstract. With the constant improvement of Chinese economic development level, there have been constantly increasing researches on artificial intelligence and database field and the application of data mining in each field has been increasingly wide. However, with the increase of information quantity and data size, such information makes it more difficult to discover effective knowledge while helping the work and production of people. A lot of information is arranged in the specified equipment. However, mode and isomerism are more complicated and network noise increases. To process such data more effectively, cloud computing method can be used to handle problems that cannot be solved by traditional distributed computation method.

Introduction

Cloud computing is a product of constant development and progress of computer, which is characterized by the ability to store, analyze and process mass data. First, cloud platform is established, where data resources are stored. Distributed computer cluster is established based on computing power, which is more stable, extensional and safe. Second, many resources on cloud platform are virtual. Such principle is open and transparent for users. Users do not need to have detailed understandings of cloud platform or professional skills. They only need to concentrate on resources used and master the method of obtaining services through cloud platform. Cloud computing can also put forward a commercial concept to improve the storage and computing abilities of computer. It can be paid on demand and actually meet user demands.

Introductions to cloud computing

Definition of cloud computing

Cloud computing is a calculation method put forward by John McCarthy in 1960s, which was first used for public utility. Service sectors provided it for users as water and electricity resources. Then, the conception of Network Computer was put forward through constant research and innovation of researchers. The setting of computer required by users in terminal was the main function of Network Computer. It could not only support users' browsing, but also allow accessing more server contents and services through browser. At that time, the basic form and function of cloud computing emerged. However, due to the limitation of extension and server terminal, network attraction and the diversity of application services were lagging and the price of PC started to decrease till the occurrence of McCarthy and Larry. Some enterprises established early promoted the development and emergence of cloud computing. Such level stayed till the occurrence of Google. From professional perspective, cloud computing can be defined as a software and hardware platform easier to be extended and utilized with IT infrastructure delivery mode and pay on demand through network. Those that can provide network resources are called as "cloud". Resources in "cloud" can be utilized and extended and their payment can be completed according to user demands. Cloud computing service in the broad sense is payment and use. Services easier to be extended are provided through network according to user demands. Such services are closely related to IT software and internet etc. and can be used for other services [1].

Features of cloud computing

Cloud computing refers to cloud storage, mainly involving the monitoring of cloud database, cloud host and cloud security, which is one of cloud service functions. In essence, it provides users with security and storage services etc. To obtain such services, computer access and the support of

remote server are required. Network is composed of multiple terminal servers, which is called as resource pool composed of server cluster. Resource usage in resource pool can be extended based on business requirements and delivery charges can be determined according to usage. Computing power circulates as a commodity. In terms of collection of charges, it is different from water and electricity. It has low costs and more convenient use. Their difference is that cloud computing service can conduct data transmission via computer. Specific features as shown as below:

Large scale: let’s take Google for example. The latest data show that Google has over 40 data centers in the globe and the quantity of servers reaches 1.2 million. There are over 51 million racks on EC2 cloud foundation platform of Amazon in the globe and the quantity of its servers reaches over 0.6 million. Cloud computing platform can establish a large resource collection and distribution center with such servers so as to improve data storage and computing abilities. Meanwhile, cloud computer is highly extensional and business can be adjusted according to user demand. For example, S2 cloud storage service has many data centers in the globe. The extension of Amazon data center will be easier according to user demand, thus increasing similar nodes. Virtualization: cloud computing can transmit many IT resources to internet with service method and finally deliver them to users for use. Storage resource, software development and system test and maintenance can be paid on demand. Cloud platform has rich resources, supports various kinds of application and has small limit on the type of application. It is also highly reliable. From technical perspective, cloud computing storage service can realize data backup and produce multiple central copies. In the case of loss of any copy, other copies can be recovered in time. For example, AmazonS2 cloud storage service can provide users with more standard storage services, the reliability of which reaches 99.9%.Cloud computing process can switch to a new node rapidly in the case of failure of a node. Moreover, technical staffs of cloud data center can manage and maintain servers in a centralized way [2].

Basic theory and algorithm of data mining

Basic theory

Data mining is a necessary process in the research and application of database technology as well as a product of constant development of database technology, which can query mass data rapidly and find out the relationship between historical data and application data. Even though data mining technology developed late, it has been widely used. Data mining refers to the process of seeking for data with potential value in database. Data mining is also a decision-making means established based on database, statistics, pattern recognition and machine learning, which can analyze enterprise data automatically, explore potential pattern and allow enterprises to establish market strategies better. In the field of IT, the source power of data mining comes from commercial field, which has intercommunity with telecommunication, finance and retail and can produce PB-level data in the short time. Business information is hidden in mass data. Therefore, data mining technology can be regarded as the process of evolution from data to information.

Classical algorithm of data mining

Making mining algorithm exist in mass data and expecting to obtain such knowledge automatically is a state that cannot be achieved. Its law can be summarized only through data mining. The pattern is generally divided into full mining of predicted and decision-making data and statistics of analytical mining task. It is established based on historical data statistics and classification. Historical law can be presented through data mining and decision-making mining task can be predicted. On this basis, new law can be found and new data set can be predicted according to the law. Both mining patterns use classical algorithm [3].

Association rule

Association rule refers to the discovery of data relationship from mass data. For example, “only 3 customers bought cucumber, but he has a probability of 60% to buy potatoes. Cucumber and potato do not have any relation, but association rule can explore such relationship. Association rule can be used in many fields. For example, banks use association rule to recommend services that users are interested in; insurance claim combination is used to eradicate fraudulent conduct; shopping basket

can be used in supermarket for product analysis and promotion and shop design and colored drawing etc. The central vocabulary of association rule is “rule”. There are many forms of rule, mainly involving “if…, ….” To judge the effectiveness of such relation, support level and confidence coefficient can be measured in the algorithm [4].

Classification and prediction

Two similar contents in mining algorithm are prediction and classification, both of which fall into different categories based on data tuple properties. They have differences in time sequence. Time plays a big role in algorithm mining. Many algorithms keep a correlation with time sequence. Therefore, prediction and judgment can be made according to conduct. Both forms require modeling. It is required to establish a data set or conceptual set in the model, constitute training sample with single tuple and then classify data tuple based on characteristic value.

Clustering analysis

Clustering analysis and sample data are similar, which can be divided into multiple categories, such as growth rate of sales volume of a transnational company in each country and clustering analysis, clustering analysis on ban card user behavior of a bank and production data of retail industry and light and textile industry. The most suitable data can be found through clustering analysis. Clustering analysis plays a bigger role in clothes buying. Data such as height, chest circumference, waistline, arm length and leg length can be obtained according to differences of body structure of males and females. Clustering analysis is a classification method, which extracts valuable information according to sample data, classifies data in a centralized way and predicts the possibility of data. Target and expected values of clustering analysis do not have significant differences in the same data set, but have significant differences in different data sets. Therefore, it is necessary to clearly organize the content and layer it based on the category.

Data mining under the support of cloud computing

Distributed computing layer: the center executing calculation and storage can support the operation of mining algorithm and solve distributed storage and computing tasks by using Map/Reduce distributed framework and cloud computing storage technology. Distributed storage service: undertake mass data storage task, provide operation time and space and have such characteristics as high stability and permanent space; parallel programming environment: execute mining task and feed back the task to application layer on the basis of Map/Reduce programming framework; node control: control of functional nodes such as operation scheduling, monitoring, distributed management and load balancing; data mining platform layer: mining algorithm layer and business layer interact, receive business command and obtain the calculation result.

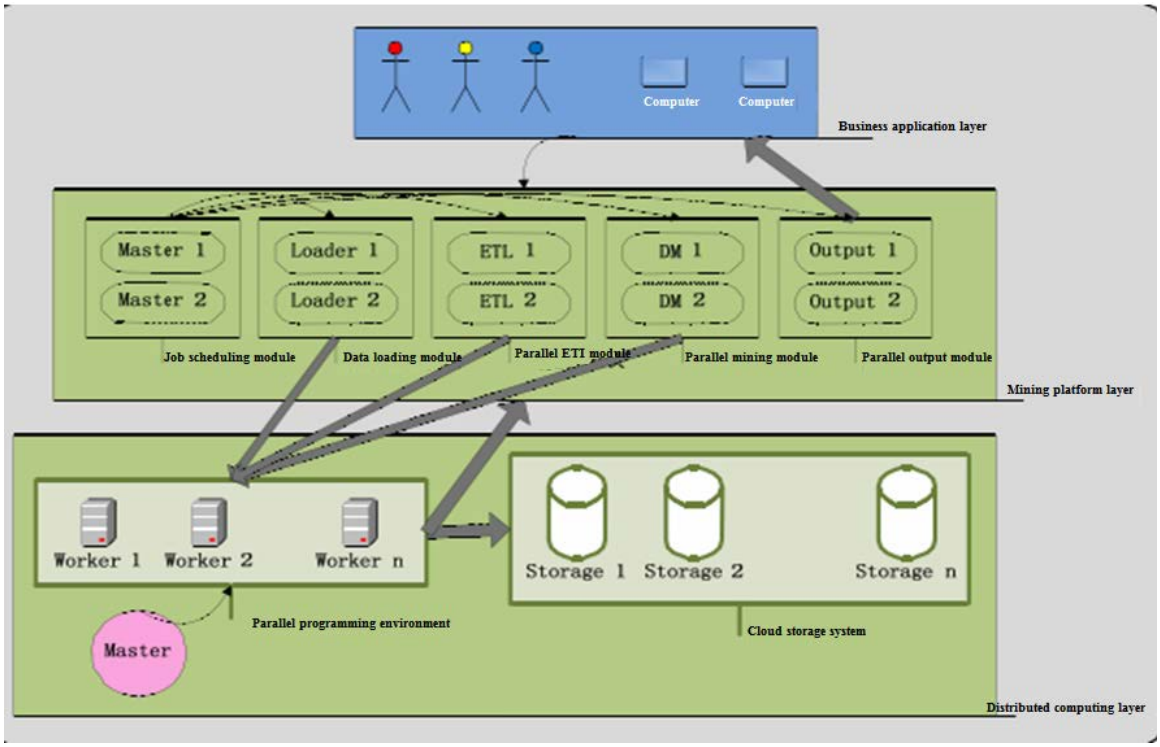

Fig.1. Framework diagram of parallel mining platform based on cloud computing

Method of data mining algorithm under the support of cloud computing

Algorithm Apriori based on Map/Reduce association rule

Any item set can be frequent item set, including many subsets. Therefore, the central theme of Apriori algorithm is that k item set is merged and extracted in k-1 item set and k item in the extracted item set is excluded. For example, records are read first in data set and a threshold value is set. 1-item subset is searched. Finally, 2-item subset can be found through the subset combination. Then, non-frequent subset can be excluded. The rest can be done in the same manner till the generation of threshold value meeting the support level set. Application steps of Apriori central algorithm are shown as below:

Single data tuple in database is calculated. Item sets 1- and C_1 can be generated. Then, minimum support level is set, which is generated from -1 and C_1 above. A total set L_1 is generated again from -1 and C_1 according to the minimum support level set. 2- and C_2 are traversed for the calculation of item set. 2- and C_2 are used as alternative item set based on the minimum support level set and L2 is generated. Repeated iteration can generate L2 and C3 and then generate L3 from C3 till the generation of frequent item set L_k. Classical Apriori algorithm follows iteration law. There are three inherent programs, i.e. database D and K- item set. Candidate K-item set can be scanned according to algorithm execution flow chart. All candidate K-item sets contain 3 data tuples. Different frequent item sets execute Join operation method, which can generate K-item set. Finally, pruning part is involved. If any subset of k-item set does not occur frequently, the item set should be removed and then the result after pruning can be obtained [5].

Clustering analysis algorithm K-Means based on Map/Reduce

Standardized K-Means algorithm is executed with the following steps: first, a clustering center is selected, e.g. cp[0]=D[0]，cp[k-1]=D[k-1]… where D refers to data of a thing. Initial center is selected randomly. For D[0]....D[n], the total number of c[i] with the shortest distance can be calculated and marked as C_i. For steps above, data tuple in D[i] has a small distance from current C[i]. However, if a threshold value is given, clustering won’t occur, algorithm program will terminate and k cluster will occur.

After Map/Reduce of K-Means algorithm, cp[0]=D[0], cp[k-1]=D[k-1] can be copied to OriginalCluster[] and subject to segmentation processing. Different computing nodes are distributed

in blocks according to requirements of node group. The distance of cp[0].....cp[n-1] is calculated respectively. The one with the shortest distance is marked as c[i] and the total number is marked as C_i. Under the framework of Map/Reduce, key-Value and Value can respectively correspond to i and D[k].

Conclusion

With the constant progress and development of cloud computing platform, a mature system has been developed and commercial computing cluster has emerged, which can be applied to EC2 of Amazon and realize monitoring, fault tolerance and scheduling. It is more reliable and safer than distributed application platform and has lower costs of use and maintenance. Therefore, cloud computing is an important means for mass data mining.

References

[1] Zou Qingchun. Study on Examination Data Mining Algorithm Based on Cloud Computing Environment. *Information Security and Technology*, 2013,4(7):18-20.

[2] Cao Xiaochun, Zeng An, Pan Dan et al. Study on Domain-oriented Data Mining Service Platform under Cloud Computing Environment. *Automation Instrument*, 2014,35(9):9-13.

[3] Ding Yan, Yang Qingping, Qian Yuming et al. Study on Framework of Data Mining Platform Based on Cloud Computing and Its Key Technologies. *ZTE Communications*, 2013,19(1):53-56,60.

[4] Chen Donglin, Fu Min, Kang Yanfang et al. Study on Data Mining-oriented Multi-case Group Purchase Decision of Cloud Resources. *Computer Application Research*, 2013,30(11):3295-3298.

[5] Fang Shaoqing, Zhou Jian, Zhang Mingxin et al. Study on Improved Selection Algorithm Based on Map/Reduce in Web Data Mining of Cloud Computing. *Computer Application Research*, 2013,30(2):377-379,395.