Research on Data Mining Algorithm and Its Application in Cloud Environment

LI Bin[1, a]
[1]Dean's office,Chongqing college of electronic engineering Chongqing 401331,China
[a]xiaohuo111@sina.com

Keywords: Data Mining Algorithm, Application, Cloud Environment

Abstract. In recent years, the rapid development of the Internet and computer-related technologies, including photography, video, e-commerce, etc., so that the data generated around us was the explosive growth, especially after the rise of the smart phone mobile Internet technology as the representative of others obvious. Faced with such a large-scale data analysis and data processing become a huge problem, which would give the opportunity to the development of data mining. Data mining can extract valuable information from users of these massive, heterogeneous, random data found interesting user mode.

Introduction

Data Mining (DM), also known as knowledge discovery in databases (Knowledge Discover in Database, KDD), since the symposium in August 1989 in Detroit, held the 11th International Joint Conference on Artificial Intelligence for the first time knowledge after the discovery (KDD) the term database and artificial intelligence has been sustained hot research question. After 20 years of development, data mining applications in various fields has been very widely. As long as a business owner in the field with the value and demand analysis database or data warehouse, such as daily global sales data, or mass entertainment resource on the iTunes Store, you can take advantage of various mining tools mining algorithm targeted of mining analysis, to discover from their massive amounts of data can be used to guide further research information or decisions. Only as an archive and appear before the application of data mining, or only so simple search queries have a new value of the mass of historical data used, one can find these historical data from the past is difficult to be understood that while manual analysis can potentially implicit information and knowledge value.

But with the development of computer technology and network technology, computers and networks are increasingly penetrated into people's daily life, study and work, we have to face today is also the mass ratio of the mass of data. And these massive data simply is not neatly arranged distributed on a limited number of devices, which are often based on a variety of complex, heterogeneous pattern, full of noisy presence information throughout the network, such as Walmart worldwide purchase, warehousing and sales data. To process and analyze the data, computing power required is often difficult to achieve in a limited time frame, and even traditional distributed cluster of computers may not be able to meet, and a dedicated network of distributed computer cluster costly, daily management and maintenance work is not easy. These problems, they gave the traditional data mining challenges.

Fortunately, cloud computing, this came into being. Cloud computing is the most recent years in the field of computer nascent a brand new revolutionary concept. Cloud computing concept natural for mass data storage, analysis and processing provides practical feasibility. First, the storage resource cloud platform, computing capabilities are based on the traditional distributed computing clusters built, easy to expand, the stability of security has been greatly improved; secondly, resource cloud platform is virtualized, its operating principle is transparent for the user, the user of cloud computing without the technical details cloud platform, do not have the appropriate knowledge and skills, they do not run the daily cloud platform to manage maintenance operations, only need to focus on their own needs use of resources and how to get to the appropriate service through the cloud platform is sufficient; the last and most important, cloud computing presents an advanced commercial concept, which makes the computer's storage capacity, computing power can be like water, traditional resource of electricity, gas, etc., as needed to apply for on-demand and pay-per-use carried out. If you want to massive commercial, financial, communications, data storage, processing and analysis, and want to see, based on data in the cloud computing platform should be consistent with a digging tool best choice.

The Definition of Cloud Computing

Under the previous business model, companies build their own computer system, you need to purchase expensive hardware amount needed equipment maintenance, software system also need to purchase a license. And, in response to business demand for equipment during the peak period, you need to configure the device to meet much higher than the usual load. Computing, storage, networking and other resources to enable them to be released into a consumable service, users only need to pay according to the amount of its actual use, this is the dream of people long since. Cloud computing is in such a case, first proposed by Google in 2007. Cloud computing will be realized as a computing infrastructure that ideal, and since all the major companies' efforts, it is fast becoming a commercial reality.

Although cloud computing, whether it is in full bloom academia and industry are discussed hot, but still lively debate on the definition painted. NIST (National Institute of Standards and Technology) is defined on cloud computing: Users can easily, on-demand access to a shared pool of computing resources (networks, servers, applications and services, storage, etc.) via a network access, with rapid deployment, minimal administrative costs of the service provider or minimal intervention of a new IT infrastructure operating mode. A popular way to understand is that cloud computing is a distributed system provided through a network of economic demarcation, the system according to the needs of users from outside an abstract, virtual, dynamic resource pool to provide users with computing power, storage, networking and other services.

The Definition of Data Mining

With the development of computer technology, we are gradually being masked data. There are a lot of data around us increases, we have done a lot of behavior will be recorded in the supermarket to buy goods selection will be recorded as electronic data, our consumer Internet, watch videos, view news will be recorded as the various logs. These records are personal choice, commercial and business areas are also similar to the same thing happening. As the data size grows, people's understanding of it is getting low. These data hidden behind a potentially useful information, but is rarely emerge or be exploited. Precisely in order to obtain these data mining useful information is presented. You can obtain the data from the data mining of their useful patterns, such as scientists can learn some rules of celestial bodies, entrepreneurs can learn from consumer propensity to consume data from a company or public astronomical data, market operation mode, etc., which do a favorable decision.

The Sources of Data Mining Algorithms Supported by Cloud Computing

With the development of information technology, the use of computers and computer networks for communication, business development, execution flow, data processing has become a part of people's daily lives must. However, a variety of complex information more and more, increasingly large scale of massive data, in order to quickly find the real and effective business can play a guiding role of knowledge in these massive data, data mining has been the classical algorithm parallelization become inevitable. In many applications have huge amounts of data, relying on small-scale distributed computing network to store, analyze data, it is still possible, but an algorithm is developed, not very easy to deploy, two distributed computing network is designed to deploy expensive , management and maintenance is also quite labor-intensive and material resources. Therefore, this chapter will examine on Map / Reduce framework Apriori, K-Means both classical mining algorithms cloud scheme, and actually be studied after the cloud computing algorithms operating efficiency, resource overhead.

Data mining technology development more than twenty years, various mining algorithms are already very well for their optimization is greatly improved their performance and efficiency. Research content of this article comes directly from Apriori, K-Means both classical algorithm and its purpose is to effectively cloud computing technology. To make the algorithm can run in the cloud platform, changes in the algorithm itself is not large, the key is how to follow Map / Reduce Framework reasonably parallelize and how to parallel algorithm after deployment in the cloud computing environment, and can be loaded cloud data storage environments operated. In fact, in the course of the development of mining technology, mining algorithm parallelization is not uncommon, but most are in stand-alone environment for the emergence of multi-threaded mode, or confined in a small distributed network. The cloud thinking under parallelism is concerned that the entire cloud computing and storage platforms in parallel on the cluster, which is the traditional algorithm modes, including the use of resources have a very different philosophy.

Data Mining Algorithm Supported by Cloud Computing

Association Rules Algorithm. Apriori association rules algorithm undoubtedly the most classic Apriori algorithm, Apriori algorithm though 1994 had already seen, but it has maintained a strong vitality in all fields has been very have a very wide range of applications. The basic idea of Apriori algorithm with other association rule algorithm is very similar, are the first to traverse through the data to find all the frequent itemsets from frequent itemsets extracted after all the rules, then the confidence is less than the preset value rules excluded. However, not all data sets can be extracted a strong rule. For example, "Haier washing machine" and "Siemens refrigerator" together, it is difficult to produce an effective rule, and later both customer brand products is actually very little. However, if the "Haier washing machine" to enhance the level of the washing machine, and "Siemens refrigerator" refrigerator level raised to be considered, this is likely to generate strong rule, because at the same time customers to buy washing machines and refrigerators is still very common.

For any set of items, if its own frequent item sets, all of which are bound to a subset of frequent itemsets. Thus the core idea of Apriori algorithm is from the frequent k-1 items centrally by combining the extracted k itemset and then there is not then the subset k itemsets frequent item sets excluded. For example, first read all the records from the data set, based on the threshold set, find frequent 1 item set and then use the frequent 1-item sets, are combined to find frequent item sets 2-, and then exclude substitution contains 2- non-frequent 1-item itemsets collection set. This continues until no longer satisfies the threshold previously set minimum support or confidence so far.

Clustering Algorithm. Data K-Means algorithm based on K as a parameter, the N data tuples in the data set is divided into K subsets, so that all the data within the subset of tuples with high similarity, and between the subsets low similarity tuples. The similarity is based on the average within a subset of the object to be calculated. The basic process of K-Means algorithm is as follows: First, all data in the data set arbitrarily selected K tuples as the initial cluster centers, other data in the data set remaining tuples, then calculate the similarity and clustering center (determined in advance according to need some kind of distance Euclidean geometry), which in turn are classified with the most similar to (on behalf of the tuple cluster center) of a cluster, obtained after re-calculated for each new merger the new data center clustering of tuples (tuple new cluster mean all data); and repeat the process until some measure functions pre-assigned (usually mean squared error function) converges to a certain threshold.

Collaborative Filtering Algorithms. With the rapid development of e-commerce, etc., as an important means of user access to information recommendation system has been widely used, and collaborative filtering algorithm is one of the most extensive and most successful recommendation algorithms used in the system, but also the field of data mining important algorithms. Users surge and the amount of product on the one hand to the recommendation system brings data recommendation required quality, it also makes megabits per second-level users and products need to be recommended. Collaborative filtering scalability and efficiency of the algorithm is facing increasing challenges. Collaborative filtering recommendation algorithm is one of the earliest and most widely used. However, the algorithm uses the process found some of the potential challenges they face, such as sparsity and scalability asked.

Not only complexity of the algorithm and the number of users and also about the number of items related. With the growth of users and projects, the algorithm often requires more computational resources or memory, this is the scalability issues. The collaborative filtering algorithm is divided into user-based collaborative filtering and collaborative filtering-based items. To solve the scalability problem, this paper proposes a run on Hadoop clusters scalable collaborative filtering algorithm based on item. Utilization characteristics Hadoop and MapReduce will calculate the amount of the division of tasks to run in parallel on different nodes.

Conclusion

The current cloud computing is a hot research and development at home and abroad, the future of cloud computing will be ubiquitous. The use of cloud computing technology, people can be distributed through a network of hardware and software integration of resources, access to computing power and storage capacity for mass data mining data providing technical solutions. In data mining, association rules analysis and cluster analysis is an important data mining algorithms.

References

[1] Huifang Zhou: Inner Mongolia Science Technology and Economy, Vol. 6 (2004) No 53, p.25-26

[2] Hongli Zhang: Science and Technology Information, Vol. 12 (2005) No 27, p.74-76

[3] Qin Guo: State Grid, Vol. 1 (2006) No 33, p.11-14

[4] Jieming Liu: China New Technology and New Products, Vol. 3 (2009) No33, p.121-124

[5] Ling Liu: Computer and Digital Engineering, Vol. 1 (2011) No 33, p.11-14