Discussion and Improvement of Apriori Algorithm of Data Mining Based on Hadoop Platform

Mengyang Zhao, Bo Tang, Le Yang

Electronics and Information Engineering Institute of Sichuan University, Chengdu, Sichuan, China, 610065

Keywords: Apriori algorithm, data mining, Hadoop platform

Abstract. Apriori algorithm can accurately find out the related items in the database, and it can be used in various fields of work and life. However, with the explosive growth of data, there are still some disadvantages in the practical application of the algorithm. This paper introduces the Hadoop platform, explores the characteristics of Apriori algorithm, and puts forward a kind of improved Apriori algorithm based on the Hadoop platform to provide some references for the relative researchers.

Concept of Hadoop platform

With the rapid growth of data on Internet, it is more and more difficult to meet the needs of massive data processing. Therefor, Hadoop is born in such an environment. Hadoop is a mainstream framework in cloud computing, it is a distributed system framework can be run on large scale clusters, which has rich experience in the development of distributed and parallel development programmers can develop distributed applications more easily, by the Apache foundation. Hadoop is a completely open source framework developed with Java, so it has a good portability. Hadoop platform has many advantages. It can reliably store and process PB - level data, and it is very simple to implement, without any modification of existing structures. It can distribute and process the data through the cluster composed of ordinary PC machine, the total number of these cluster nodes can reach tens of thousands. Efficient data exchange of distributed file system and MapReduce parallel processing of data, making the processing more quickly. The redundant backup mechanism of HDFS and the task monitoring mechanism of MapReduceduce make it possible to re deploy the computing task after the failure of the mission, and to ensure the reliability of the system. Hadoop platform is a popular and important framework of cloud computing, which is a distributed system infrastructure developed by Apache foundation. Because when using Hadoop, the user cannot understand and master too much developed in parallel and distributed development experience, the Hadoop platform has many developers, and in a short time after the launch will attract a lot of people to use and Study on its platform, let the rapid promotion. Hadoop is a software framework that is reliable, efficient, scalable, and able to distribute large amounts of data. Because it works in a parallel manner, you can use JAVA technology to deal with the data of PB level on the cheap PC machine. The Hadoop platform can be used freely.

Relevant theory of Apriori data mining algorithm

The Apriori algorithm is an association rule mining algorithm proposed by Rakesh Agrawal and Ramakrishnan Srikant in 1994. It was first proposed in connection with the pruning technique based on support, control the growth of candidate itemsets. The purpose of association rule mining is to find out the relationship between data items and items. Living in an era of information explosion, all around us are surrounded by all kinds of information, how to obtain useful knowledge for us in the ocean of information, data mining is one of the best ways, it can let us get the most valuable information from vast and complex information. Therefore, data mining has been widely concerned by scholars, the algorithm has been widely used in various fields, has become the most influential algorithm. Data mining is to extract the hidden information from the fuzzy, noisy, random, a large amount of information, hidden in it, potentially useful information. In simple terms, is to explore the information needed from the database. Data mining is widely used in many fields such as artificial intelligence and robotics. As the name suggests is to find out the correlation of association rules from large amounts of data, Agrawal first proposed the equivalent to the problem of association rules mining customer transaction databases, analysis is the most typical shopping basket, they carry on the optimization to the original algorithm, such as random sampling method to improve the efficiency of data mining and data mining application association. Apriori algorithm is one of the most influential algorithms for mining association rules. The core of the algorithm is to generate frequent item sets by two stages, which are candidate set generation and down detection.

Necessity of the improvement of Apriori data mining algorithm

Apriori also has very obvious shortcomings. Apriori algorithm may produce a large number of meaningless or even false candidate sets, and there are a large number of redundant data between these candidate sets, resulting in low efficiency. Although the Apriori data mining algorithm provides a lot of help for the complex problem of association rules mining, but the traditional Apriori data mining algorithm still has some shortcomings and deficiencies, here are some of the shortcomings of Apriori algorithm of data mining. Apriori has two major steps. First, to find all the large frequency set. The frequent occurrence of these item sets is the same as the predefined minimum support. Then the strong association rules are generated from the frequency set. Rules must satisfy minimum support and minimum confidence. In the Apriori data mining algorithm, we need to do a lot of comparison of the previous items. The pruning step algorithm in mining Apriori data. When the scale of the database is large, it is a time-consuming work to scan the database, and the efficiency of the Apriori data mining algorithm is not affected by the multiple scan databases. Traditional Apriori data mining algorithms can only deal with a single transaction data set. Since the Apriori algorithm is based on the idea of layer by layer search, it is necessary to scan the database again to determine which items are frequent. This method increases the burden of input, output and is very time-consuming. It cannot be directly used for association rules mining in relational databases. It cannot directly deal with numerical data. The model describes the frequent

relationship between the projects. The purpose of the status is equal, but in fact not the case. When the traditional Apriori data mining algorithm is too large in the database, due to the limited processing capacity of the single machine, it needs to keep the data set in and out of memory. The system load is too large, resulting in low efficiency.

Paths of the improvement of Apriori data mining algorithm based on Hadoop platform

The main process in the process of reading the original data set on the set of statistical comparison produced the first candidate. For the candidate, the author uses the Map Hadoop function under the framework of the original data set is divided into several sub sets, then distribute parallel to the Reduce function, through the support of statistics and screened the first frequent item sets, followed by first order frequent item sets generated second order candidate. And so on, until the K order candidate. The database is divided into horizontal size n disjoint blocks of data, and then sent to the M node; each node scanning data blocks respectively, generate local frequent k- item sets in each block, the local frequent item sets generated 1- process is similar to the frequency statistics, frequent k- set the serial with traditional Apriori algorithm. All local frequent k- item sets are merged to form global candidate frequent item sets. Frequent item sets local frequent item sets may not be D, but any D frequent item sets as local frequent item sets and at least one part again scan D, calculated for each candidate frequent item sets support threshold and minimum support threshold comparison, determine the final set of frequent items. It only needs to scan two times D to mine all the frequent item sets, which reduces the overhead of the algorithm execution. Each node local frequent item sets mining process is not dependent, it can reduce the communication among the nodes to improve the efficiency of the algorithm. All individual items are candidate item sets. Any support value smaller than the given minimum support value will be removed from the candidate set to form a frequent item set. Determine the support of these options by scanning the database again. The candidate items with large minimum support degree are given to form frequent item sets.

The era of big data, Apriori algorithm has some shortcomings: Apriori algorithm need to scan the database in the process of implementation, the data set size under the condition of a large data set in each scan for a very long time, resulting in the efficiency of the algorithm is very low, even failed to run. An effective strategy to solve the problem of the traditional Apriori algorithm is applied directly to the Apriori algorithm and Hadoop platform combining the large data set in accordance with the Hadoop way into small pieces of data, and distributed to different computing nodes, greatly reducing the amount of calculation on a single computer, the algorithm for parallel execution. Hadoop platform can provide a variety of complex algorithms to achieve parallel conditions, but also for large-scale data storage security. So, we will MapReduce framework of the improved Apriori algorithm is transplanted to the Hadoop platform, the implementation of distributed parallel processing, so as to reduce the traditional Apriori algorithm in the single on the pressure and improve the efficiency of the algorithm Apriori. In order to K+l data, if required for each order K+l in matching K according to conventional methods, but according to the previous method, only need to order K frequent item sets to generate a subset of read at the end of the module can achieve the desired purpose, then. Read frequent item sets. In this process, both the time

dimension and the spatial dimension, the complexity can be greatly reduced. According to the inherent characteristics of the database, which is divided into several areas unrelated, here the partition size to make blocks can be summed up to the total score in the area, are scanned only once, then consider all frequent sets generated by each region individually, then merge all frequent sets. We use all possible association rules to calculate the support of these frequency sets. Each region can be assigned to a processor to generate a frequency set. After all the frequency sets are processed, the processor can communicate to generate a global candidate frequency set. This algorithm can improve the correct rate of association rules to eliminate a large number of meaningless repetitive data.

In this paper, some candidate itemsets C_k in the improved algorithm are generated by C_{k-1}. C_{k-1} scale is generally larger, resulting in the number of times required to connect a lot. In the set are sorted by default dictionary order case, when C_{k-1} and L_1 connection, we only need to compare the size of C_{k-1} and the relationship between the last one, to avoid the self-connection required for a lot more things happen. In life, different projects have different uses. It is necessary to mine the hidden information hidden in the database and obtain its value. For example, department stores, it needs to know what consumers need in different seasons, different ages, so as to be able to have a tendency to recommend and promote goods, get more profits. In order to find out the importance of different projects, the need for different weights on the set, the project is divided into different categories according to the logic, the weighted value between different categories, setting the minimum weighted support value, using the Apriori algorithm, calculate the weighted support, to meet the candidate set weighted value for screening in addition, finally can dig out the inherent correlation value according to practical needs. Compared with the original algorithm, the improved Apriori algorithm has been improved in many aspects. The improved data mining algorithm saves a lot of time, because it takes only a scan of the database, in the data mining process does not need to scan the database to determine the support of itemsets, solves the shortcomings of the original repeatedly scanning the database in the Apriori algorithm, to get the frequency itemsets with minimum intersection support to the rest of the calculation can be done in memory. The object of the improved Apriori algorithm will be clear, other items will be involved in the combination of processing each generated item set, delete some do not meet the requirements of the item set, reduce the space complexity, which sets clear objectives, so as to improve the efficiency. Due to the fact that there is no parallel scanning on the database, the database scanning time is still not changed, and the time consuming is relatively more. When the algorithm is transplanted to the platform in parallel, because the database is not mean to different computing nodes, each node scans the database is smaller, the time of scanning the database can be reduced a lot, it can be said to some extent reduce the running time of the algorithm, improve the operating efficiency of the algorithm.

Conclusion

Based on the analysis of the original Apriori algorithm and the characteristics of the Hadoop platform, the algorithm is transplanted to the core of the Hadoop platform, and the algorithm is further improved. The improved algorithm of Apriori has a significant improvement in the

operation rate. Next, we plan to further combine the Apriori data mining algorithm Hadoop in the larger cluster test to try to improve the algorithm.

References

[1] Sun Zhaoxu, Xie Xiaolan, Zhou Guoqing, Ni Jinsheng, Hu Xin, Research on Apriori algorithm and implementation of Hadoop platform, Journal of Guilin University of Technology, 34(3), pp. 584-588, 2014.

[2] Zhu Jintan, Improve of data mining Apriori algorithm, Electronic Design Engineering, 21(15), pp. 37-40, 2013.

[3] Wang Yingbo, Ma Jing, Chai Jiajia, Zhao Bin, Improved Association Rule Mining Algorithm Based on Hadoop Platform, Computer Engineering, 42(10), pp. 69-74+79, 2016.

[4] Chen Zhigao, An Improved Ant Algorithm QoS Routing on Hadoop Platform, Fire Control Radar Technology, 43(1), pp. 22-25+55, 2014.