Research on Multidimensional Association Rules Algorithm based on Hadoop

Yuanyuan CHENG
School of Information Sciences and Engineering, Chongqing Jiaotong University, Chongqing, 400074, CHINA

Abstract: The existing parallel multidimensional association rules algorithm has a lot of problems. On account of the data large and disorder makes huge communication traffic, no uniform distribution cannot deal with load balancing, it is also makes the system I/O performance is low, the new parallel multidimensional association rules algorithm based on the Hadoop platform is proposed. Thought of the new design algorithm adopting the method of equip-width discretization to split each attribute domain, after a scan, deleting the dissatisfaction property value of each box, then mapped to a unified space. After preprocessing, combining the improved Apriority algorithm for implementation by using the Map Reduce programming model. The results show that the improved algorithm can better solve the above problems on the basis of improving the efficiency of mining, and this algorithm has great value in research and utilization.

Keywords: Multidimensional association rules algorithm; Huge communication traffic; Load balancing; Equip-width; Preprocessing

1. Introduction

Nowadays, along with the rapid development of network technology, mobile internet technology and social networks, data size and the type of data is growing at an alarming rate, the big data era has arrived. According to IDC predicts that by 2020, the volume of the whole world will reach 35.2ZB.But in addition to the large amount of data, the big data also including the diversity and complexity of the data, and complexity including the multidimensional data, such as the time dimension, space dimension and so on. The multidimensional nature of the data increase the greater challenge on the basis of the huge amount of data that is not easy to deal in is difficult to complete the task of the traditional machine learning algorithm in the big data era. With the continuous development of distributed processing technology [1], Hadoop is more and more popular. The Map Reduce distributed computing model provided by Hadoop can guarantee the computing power, data security and data reliability of massive data.

At present, the use of machine learning algorithms to solve large-scale data set of research has been a lot of kinds. Generally use MPI parallel, PVM parallel, or based on CPU, GPU, high performance cluster in former algorithm to write more complicated. The latter relatively high requires for hardware is not suitable for large-scale ccccc the distributed programming model(ccc) under open source software Hadoop[2] fully make up for the deficiency of the c cc processing mass data has two advantages:1)the users realize the function of the algorithm only need to write a map function and reduce function;2)users don't need to care about the details of data storage, distribution, copy and load balancing under the distributed c using the advantages of Map Reduce has achieved a lot parallelization algorithm on the MapReduce. The purpose of this paper is to research correlation algorithm in data mining by using c -ccc c the algorithm has been relatively mature, but still has some limitations to process massive amounts of date which is multidimensional. And at present the research of parallel multidimentional association rules algorithm is relatively small. New design algorithm process multidimensional by reasonable dividing method. Then greatly improve the performance of mining on the parallel algorithm.

2. The Related Theory

2.1. The Programming of Hadoop and Map Reduce

Hadoop is an open source under the Apache software, which is a distributed System infrastructure. Providing distributed storage system(HDFS) and distributed programming model(Map Reduce) and distributed database(Base) for huge amounts of data. HDFS characteristics such as high fault tolerance and high scalability allows users deploying Hadoop on the cheap computer hardware, forming a distributed system; MapReduce model is a kind of parallel programming model which implementing on the basis of HDFS; HBase is an open source, distributed database based on the storage model, suitable for storage of unstructured data. In addition to Hive, Pig, ZooKeeper, Cascading model, etc.

Map Reduce adopting the idea of "divide and rule', a common explanation is "task decomposition and results of the merger". Users only need to write a map function and a reduce function. Then realizing the function algo-

rithm by using of an input<key, value>collection to produce an output of<key, value>collection. And allows the user to develop parallel applications in a distributed system that did not understand the underlying details.

Map and Reduce phase process of data processing is shown in Figure 1.

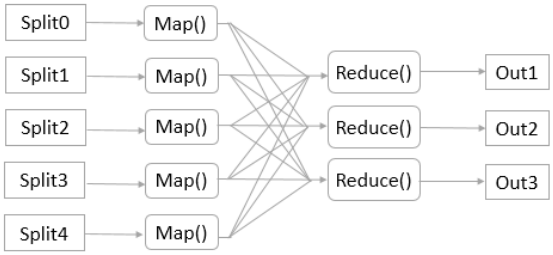

Figure 1. Map Reduce programming model

2.2. Multidimensional Association Rule Mining

Association Rules is an important branch of data mining technology research, and most of the classical algorithms from Association Rules are Apriority interpretation and improvements. The purpose of the Association Rules is to find the min_sup and micron which can meet user-specified.

(1) Apriority Algorithms

The first proposed Apriority algorithm by Professor Agrawal in 1993.It aims to dig the correlation and dependencies of massive data between things. Then many scholars and agencies have done a lot of research for Association Rule. As many improved algorithm, quantitative association rule mining algorithms, parallel mining algorithms and incremental mining algorithms based on the typical Apriority Association Rule. Apriori algorithm mainly adopted the way of layer by layer, obtaining k+1 item sets from k items candidate. Specific comprises two main process steps: connections and pruning.

Apriority algorithm structure is simple, and easy to understand and implement. However, in order to obtain frequent itemsets, Apriori algorithm need to scan data sets many times. This process will take a lot of time and storage space. If processing large databases by using traditional Apriority algorithm, computation and I/O capacity will be very large. The data of large databases typically reached TB-level, even PB-level, it is clear that serial algorithm [3] cannot meet the requirements. So the study of parallel algorithms is a must.

(2) The Outline of Multidimensional Association Rules

However, with the advent of the Internet information age, not only the amount of data continues to surge, data types, attributes are also more numerous. Data processing is the trend of multi-dimensional development of the information society. So the research of multidimensional data parallel association rules are popular research direction in the field of data mining.

According to the rules of data mining, dimension data can be divided one-dimensional Association rules and multidimensional Association rules.

In the multi-dimensional association rules, using a more efficient data processing model to process multi-dimensional, distributed processing large data sets have been generated a great deal of influence. So before making association rules, the first is data preprocessing. As you can improve the efficiency of data mining by specifies the user interested dimension and effective division of data dimensions. Pretreatment of multidimensional reasonable not only can reduce the data mining process, but the excavated multi-dimensional association rules will be more rationality. And the association results also will be more meaningful.

3. The Parallelization Improvements of Map Reduce

Big data era, filter the data, process, analyze and extract useful information by using the form of cloud computing has become important research ereas. And now generally the data is multidimensional and huge number. If using the existing parallel algorithms to analyst such data, not only will spend a lot of time and storage space, while the computation and I/O capacity is very large, and the result is also complex and diverse. So it is not conducive more effectively valuable information for people. To solve this problem, in this paper, a parallel as the goal. firstly, having effective treatment for multidimensional data, and then combined with improved parallelization Apriority algorithm, finally proposed MD_Apriori algorithm. The algorithm improved efficiency and reduced the number of scans of the data file, also solving the problem of load balancing, and greatly reduce the I/O load on the system.

First, in order to better achieve the data mining of multi-dimensional association rules, the need for multidimensional data preprocessing is necessary. The quality of data preprocessing directly related to the reasonable of data mining results. Therefore it is important to have a conversion for multidimensional data.

Assumed that the data set is D,D includes m a n-dimensional data objects, then data samples are denoted as Where 1≤i≤m.According to all the data samples, we can obtained the range of each dimension attribute is: $V_j = \left[\min Value, \max Value\right]$, 1≤j≤n.

To better achieve parallelism on Hadoop platform, the range of each attribute are divided into M parts. Then getting M range blocks. Each dimension attribute that is divided into the same width as M box: $V_{jl} = \left[\min Value + \frac{|V_j|}{M} \times l, \min Value + \frac{|V_j|}{M} \times (l+1)\right]$, (1≤j≤n, 0≤l≤M).

Then all the data blocks are sent to M node, obtaining a set of frequently set L1 of each block by using Map Reduce computation model. According to the set of minimum support, you can delete the property value which is less than the minimum support. So during the relevant previous algorithm, we have eliminated a large number of non -conforming or abnormal data. This not only reduces the storage memory, but also reduces the data scanning time and calculation when executing association rules.

Establishing the coordinate's press box, using a row vector rid to response the attribute infarction of data samples. The row vector is represented as: $Rvid = \left(X_1, \ldots, X_n\right)$, $X_i \left(1 \le i \le n\right)$, is the number range blocks the itch attribute belongs, Number in the range [1,M].

The row vector Rid is represented the unique identifier of coordinate divided blocks. Each piece contains a sample data set recorded as BlockRvid, it can be expressed as:

$$Block_{Rvid} = \left\{ T_i\left(I_1, \ldots, I_n\right) \middle| \begin{array}{l} I_j \in V_{jX_i} \\ 1 \le j \le n \\ 1 \le i \le m \end{array} \right\} \Rightarrow \begin{pmatrix} split1 \\ split2 \\ \vdots \\ splitM \end{pmatrix}.$$

Because it is m an n-dimensional data, each dimension attribute are divided into M blocks, so a total of MN blocks BlockRvid. The MN blocks BlockRvid average assigned to a computer node M.Th. data set is divided into M parts which are similar size.

From the above MD-Apriority, the algorithm only needs to scan the database twice. Adopting the method of equip-width discretization to split each attribute domain, after a scan, compression of things. When executing association rules need to scan the database once. So this algorithm overcome the bottleneck that Apriority algorithm require multiple scans database.

4. The Analysis of Experiments and Results

4.1. Experiment Environment

In this study, we selected four ordinary PC which system using Ubuntu Linux 12.04 to build Hadoop cluster platform. Adopting Map Reduce programming model and Java language to realize algorithm. One of which serves as the Master node, the other three as the Slave node. Four PC's hardware configuration and the basic setting are shown in Table 1.

Table 1. The configuring of PC

Hardware environment	
CPU	Intel®CoreTM i3 CPU 530@2.93GHz
RAM	2G DDR2
Storage	500GB SATA
Network	100M Fast Ethernet
Software environment	
Hadoop	Hadoop 2.6.0
JDK	JDK 1.6.0_21

4.2. Data Collection and Results

This experiment data used Chongqing monitoring data from a bridge. According to the characteristics of the data set collected, dividing into three types of data sets.

Strain - sectional data set A:9-dimensional data composed and size of 12G;

Vertical displacement data set B:20-dimensional data composed and size of 8G;

Solar power data sets C: 8-dimensional data composed and size of 4G.

To verify the improved MD_Apriori algorithm in the cloud computing model, this paper adopting the above-mentioned three data sets to have an experiments under different environments.

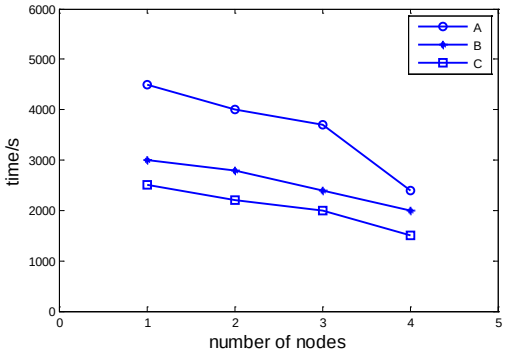

Figure 2. The Test Results of different data sets

As can be seen from the experimental results, as the number of nodes increases, the same data set the test time constantly shorten. And when the data set is increasing, with the number of nodes increases, the test time becomes faster decline continues, the performance of this algorithm reflect the more obvious.

5. Conclusions

Aiming at the multidimensional association rules algorithm based on hadoop, Firstly, it introduces the Hadoop framework platform, MapReduce programming model, the traditional Apriority algorithm and multidimensional association rules algorithm. Then adopting the method of equip-width discretization to split each attribute domain, after a scan, deleting the dissatisfaction property value of each box, then mapped to a unified space. After preprocessing [4], combining the improved Apriority algorithm for implementation by using the Map Reduce programming model. The results show that the improved algorithm can better solve the load balancing on the basis of improving the efficiency of mining, and also can deal with the problems of lower efficiency because of the database I/O operations excessive. This algorithm has a wide range of research and application.

References

[1] EUI-HONG HNA, GEORGE KARYIS. Scalable Parallel data mining for association rules [J]. Knowledge and Data Engineering, 2000, 12(3): 327-341.

[2] ROBERO, BAYARDO. Mining the most interesting association rules in knowledge databases [J]. Machine intelligent, 1999, 15: 145-154.

[3] Wegener D, Mock M, Adrenal D, etal.Toolkits based high-performance data mining of large data on Map Reduce clusters[C] // IEEE International Conference on Data Mining Workshops. 2009: 296-301.

[4] Yu Kunming, Zhou Jay, Hong TP, teal. A Load-Balanced Distributed Parallel Mining Algorithm. Expert Systems with Applications, 2010, 37(3): 2459-2464.