Association Analysis of Large Sample Data Based on Hadoop

Ran An[1,a], Jingchang Pan[2,b]*
[1,2] Shandong University, Weihai, 264209, China
[a]email: anran0708@163.com,[b]email: jingchangpan@163.com

Keywords: Hadoop; Mahout; Association Rule Mining; FP-growth Algorithm; Pattern Assessment

Abstract. This paper implemented effective associate rule mining based on Hadoop parallel computing. First, the parallel FP-growth algorithm was run on Hadoop platform to find the frequent item sets of the transaction data. Second, the strong association rules was generated from the frequent item sets by a designed algorithm. Then, redundant rules were deleted according to filtering conditions to make model evaluation. After those steps, all the funny and non-redundancy strong association rules were mined out. In addition, this paper also analyzed the efficiency of the Hadoop parallel computing and explained the superiority of the Hadoop parallel computing when it handles big data.

1 . Hadoop Platform

Hadoop is an open source distributed computing platform[1] using a lot of inexpensive computers to build a Hadoop cluster. Through deploying applications to every computer of the Hadoop cluster to implement the parallel processing, speed up the processing speed and extend the storage space. Its superiority is to handle of large data in parallel. In addition, Hadoop has a lot of merits, such as high reliability, high scalability, high efficiency, and high fault tolerance and so on.

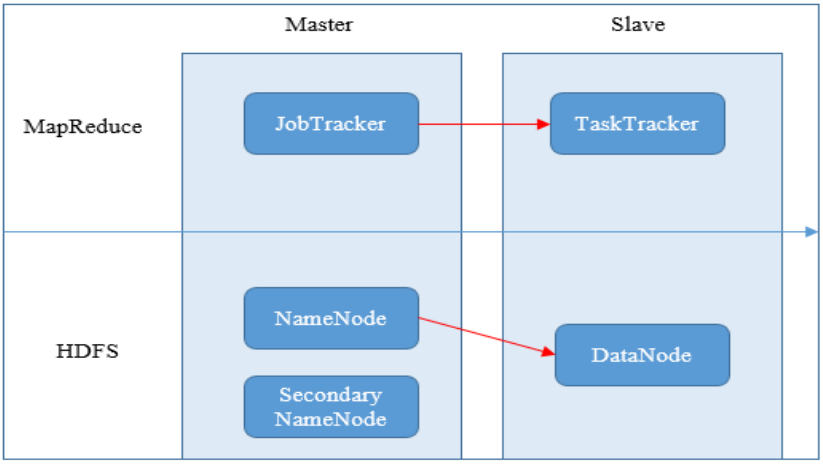


Fig 1. Hadoop's architecture

Hadoop's architecture is shown in figure 1. The master node includes JobTracker, NameNode, and Secondary NameNode. The slave node includes TaskTrackerand DataNode[2]. And the master node manages and assigns tasks, while the slave node executes the task. The cores of Hadoop are HDFS and MapReduce programming model. The HDFS file system is used to deal with lager files. It adopts stream access and cannot be modified, only can be deleted after written in. the HDFS file system consists of NameNode, Secondary NameNode, DataNode. The NameNode is primarily responsible for managing the cluster's namespace and configuration information, the DataNode is used to store data. As a distributed programming framework, MapReduce takes the idea of "divide and conquer" through compiling MapReduce programs[3].Tasks are assigned to multiple computers to process in parallel.

2.Association Rule Analysis

2.1 FP-Growth Algorithm

As an excellent association mining algorithm, FP-growth algorithm[4]doesn't need to generate large numbers of candidate formulas. It produces a frequent pattern tree (FP tree [5]) and then finds out all the frequent item sets based on the tree. According to a series of studies, FP-growth has better scalability and effectiveness than Apriori algorithm.Mahout[6] is a software package that contains a set of well-known algorithms, including FP-growth, running on Hadoop platform using MapReduce programming framework. By using Mahout's commands, we can generate the frequent item sets simply.

2.2 The Clip of Association Rule

If using two metrics like Min-Support and Min-Confidenceto generate strong-association rules, it will most likely occur that multiple rules express the same meaning and get the same conclusion, have semantic repetition. Some of the rules are called redundant rule. But those redundant rules can be avoided. Here are several common forms of redundant rules:

(1) If rule A->B and rule C->D meet the conditions like $A_{\cup}$ B=$C_{\cup}$ D=Ik and $A_{\in}$ B, then rule C->D is called as rule A->B's simple redundant rule.

(2) For rule A->B and rule C->D, $A_{\cup}$ B=I1 and $A_{\cup}$ B=I2, if $I2_{\in}$ I1 and $A_{\in}$ C, then rule C->D is called as rule A->B's strict redundant rule.

(3) For rule A->B and rule C->D, if rule A->B can push out rule A->C, then rule A->C is called as redundant rule [7].

2.3 Pattern Evaluation Method

After a series of steps, the strong-association rules we get are not all intersecting and don't conform to the practical significance, which also contains some negative correlation and boring strong-association rules, so we have to make a series of pattern evaluations[8]. This article uses a metric which is called Lift to make pattern evaluation. Through calculatingeach rules' value of Lift to select all of the intersecting rules. For rule A->B, the Lift's calculation formula is as follows:

(1)If the Lift<1, A and B are negative correlation. So one happens may lead to another does not appear. We need to delete the rule.

(2) If the Lift=1, A is independent of B. So they have no correlation and the rule is not an intersecting rule. We need to delete the rule.

(3) If the lift>1, A and B are positive correlation. So the rule is intersecting. We keep the rule.

3 . The Association Analysis Experiment of Large Sample Data Based on Hadoop

3.1 Experiment Environment

The experiment carries out on Hadoop platform. The Hadoop cluster consists of nine computers, oneas Master node, and the others as Slave nodes. The version of the operating system is fedora20, Hadoop's version is Hadoop2.20, and the version of the Mahout is Mahout0.8. The development tool is Eclipse.

3.2 Experiment Data

The experiment data comes from part of USA census data (1994). The census data has 11 columns, those columns are Age, Workclass, Education, Marital-status, Occupation, Relationship Race, Sex, Hours-per-week, native-country, Salary[9].

3.3 Experiment Content and Steps

Step1. Generate the frequent item sets:In this step, using Mahout built-in FP-Growth algorithm to generate the frequent item sets.

Step2. Generate the strong-association rules:According to the frequent item sets obtained in the first step, we write programs to generate the whole rules and calculate every rule's value of Support, Confidence and Lift. If the rule's value of Support and Confidence meet the value of Min-Support and Min-Confidence, we consider it as a strong-association rule.

Step3. Model Evaluation of Association Rules:In this step, we need to design algorithm to calculate the Lift valueof each rule and screen out the rule with Lift value being greater than 1.

Step4.The Trim of Rules: Using the redundant conditions to delete all of the redundant rules.

Step 5. The Performance Analysis of Hadoop:Analyze the running timeof different task size on different cluster node number by altering the configuration files of the Hadoop.

3.4 Experiment Results Analysis

The Performance Analysis of Hadoop

The Figure 2 shows the running times when the tasks' data size are 6.8M, 204M and 2.2G and the number of the Hadoop cluster's nodes are 1,3,5 and 8. The running time of each run are different, because the statuses of Hadoop are different, so the data shown in figure 3 are the average value of each run. The figure shows that the running time for small samples doesn't decrease obviously with the increase of the cluster's node number. However, the running time for large samples decreases obviously with increase of the cluster's node number. The reason for this is that the master node needs to spend times to distribute work to every slave nodes, when the experimental data are relatively large, the time spent to distribute work can be ignored. In this case, the advantage of the cluster will be reflected obviously. So we can draw a conclusion that Hadoop has it advantage when dealing with large samples over small samples.

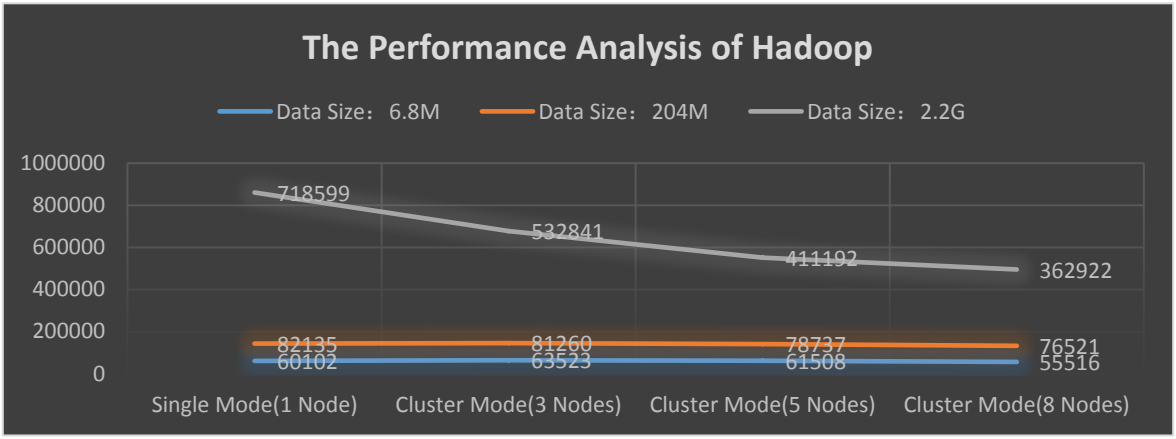
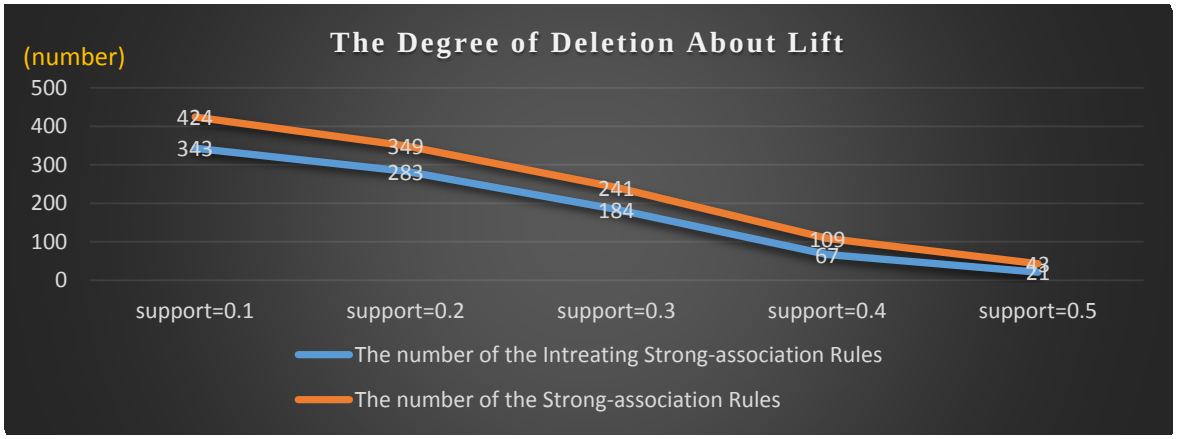


Fig 2.The Performance Analysis of Hadoop

Fig 3.The degree of Lift's deletion when the value of Min-support is 0.7



The Analysis of The Degree of Pattern Evaluation's Deletion

When the value of the Min-confidence is set to 0.7 and the Min-support is set to a series of different values, the Figure 3 shows the relationship between the Interesting Association Rules and the total Association Rules after deleting the Tedious Association Rules in pattern evaluation step. We found that using the Lift to make pattern evaluation will reduce the number of the useless rules greatly.

Acknowledgement

In this paper was sponsored by The Natural Science Foundation of China (Project No.U1431102). *Corresponding author: Jingchang Pan.

References

[1] LuJiaheng. Hadoop in Action. Beijing. Machinery Industry Press. 2011.

[2] Tom White. Hadoop: The Definitive Guide. Tsinghua University Press, 2010.

[3] XieYaowei. Savor Hadoop-Hadoop Cluster (Section 8). Shrimp studio. 2012.

[4] Han Jiawei, MichelineKamber, Jian Pei. Data Mining: Concepts and Techniques. Beijing: Machinery Industry Press.2012.

[5] Zhou Shihui. The Research and Application of an Improved Parallel FP-Growth Algorithm Based on Hadoop: [D]. Shandong University .2013.

[6] Sean Owen, Robin Anil, Ted Dunning, Ellen Friedman. Mahout in Action. Manning Publications.2010.

[7] Zhang Xinxia. The Interesting Association Rule Mining Based on Statistical Correlation: [D]. Wuhan University of Science and Technology. 2002.

[8] Ye Fuming. Basic Research on Data Mining Technology. Manufacturing Automation .2011.