The New Algorithm of the Item-based on MapReduce

ZHAO Wei[1, a]

[1] College software Technology School, Zhengzhou University Zhengzhou 450002 ,China

[a]iezhaowei@163.com

Keywords: Recommendation system parallel computing Clustering

Abstract. Traditional collaborative filtering algorithm based on item and K-means clustering algorithm are studied, the parallel algorithm of collaborative filtering Item-based on MapReduce is proposed by using MapReduce programming model. The algorithm is mainly divided into two steps, one step is K-Means algorithm clustering for users, another step is the parallel Item-based algorithm for clustering user recommendation. Experimental results show that the algorithm has obtained very good effect, improved the running speed and execution efficiency, the improved algorithm is much suitable for processing big data.

Introduction

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization. Volume means big data doesn't sample; it just observes and tracks what happens; Velocity means big data is often available in real-time; Variety means big data draws from text, images, audio, video; plus it completes missing pieces through data fusion[1]. Therefore, the big data must be through the computer statistics, comparison, analysis of the data can be the objective results. Now electronic commerce systems of every transaction, every input and every search can as data, data through the computer system to do the screening, sorting, analysis, so that the analysis results is not only an objective conclusion, more able to help business provided the decision-making of enterprises and also collected useful data can also be reasonable planning, actively guide the development of larger power consumption, and more effective marketing and promotion. With the increasing amount of data in the electronic commerce system, the need for a large number of data depth analysis is increasingly urgent. Therefore, the use of a simple and high scalability of the program for the analysis of product recommendation is particularly important. At present domestic many ecommerce sites use collaborative filtering algorithm, such as Amazon, Dangdang, collaborative filtering algorithm is mainly divided into based on the items of the collaborative filtering algorithm and user based collaborative filtering algorithm. Based on items of collaborative filtering algorithm is to measure the similarity between items according to the user's preferences, do not need to consider the item specific content features, so the algorithm is mainly used in e-commerce recommendation and movie recommendation domain, the algorithm while in the field of electronic commerce recommendation has been a certain degree of success. But in massive data are recommended when the data is recommended performance is not high and the data information lack of sharing and extended the lead to the hardware requirements compared higher inherent shortcomings make it did not receive a promotion and support of enterprise electronic commerce[2]. So if we use MapReduce to achieve distributed parallel computing, it will greatly improve the efficiency and performance of the algorithm, and promote the further development of the algorithm[3-4]. Based on the items of the collaborative filtering algorithm is according to item similarity and user history access record recommended to the user to generate a list of items, but there are some small problems, such as data sparsity problem and when the mass of users and the number of items, the user behavior and record data will greatly, and the algorithm for

computing items with similar matrix cost greatly, algorithm efficiency and performance will greatly reduce. Aiming at the above problems, the clustering algorithm has also been applied to a collaborative filtering algorithm based on item, the massive user clustering analysis, so it can avoid the question carefully, for each user to recommend operation. The first shopping users with similar interests into a user class, with a cluster of user recommended goods are the same. The second is to reduce the massive user dimensions become dozens of clustering limited, the time complexity encountered a bottleneck, and the parallel clustering algorithm using MapReduce is the effective way to solve the bottleneck[5]. MapReduce is a distributed programming model framework on Hadoop platform, in the condition of not familiar with the underlying details of the distributed implementation of the implementation of the program [6]. The MapReduce as parallel computing programming model, first of all to users of MapReduce based parallel clustering and according to the results of user clustering, in every user class using the MapReduce parallel collaborative filtering recommendation, eventually give users a reasonable personalized commodity recommendation list. The running time of different nodes in the quantitative data is compared with the new algorithm. The results show that the data processing performance of the proposed algorithm is greatly improved.

The principle of MapReduce programming model

MapReduce is in Hadoop platform by using parallel computing programming model, the technique is proposed by Google for a typical distributed parallel programming model, the user in the MapReduce model develop the map and reduce functions, can realize the parallel processing. Map will be responsible for data dispersion, Reduce is responsible for data aggregation. Users only need to achieve Map and Reduce two interface, you can complete the calculation of TB level data. Because of the MapReduce model, the details of the parallel and fault-tolerant processing are encapsulated, which makes programming very easy to implement. MapReduce parallel calculation is divided into two parts, the first step is initializing the original input data file and the data set is divided into a plurality of a certain size of data block, facilitate parallel computing; the second step is to start the map and reduce functions algorithm of parallel computing, finally produced the final result.

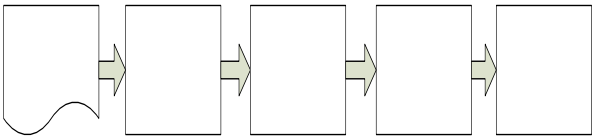

Figure1 Parallel flow chart of MapReduce

Key technology research and Implementation

1. The basic idea of the traditional collaborative filtering algorithm based on Item-based

The traditional based on items of collaborative filtering algorithm the basic idea is divided into three parts, the first part is to compute the similarity between items, common similarity calculation method with cosine similarity, Pearson correlation coefficient, Tan moto coefficient correlation of. This paper selects the Euclidean similarity algorithm, as follows:

The assumption is that there is a vector X and a vector Y： X=(x_1 , x_2 , x_3),Y=(y_1 , y_2 , y_3), Using the Euclidean similarity algorithm to calculate the similarity between X and Y S vector (x, y) formula is as follows[7]:

$$S(x,y)=\frac{1}{1+d(x,y)} \qquad (1)$$

Where $d(x,y)$ is the distance between the vector X and Y, the calculation formula is as follows:

$$d(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+(x_3-y_3)^2} \qquad (2)$$

The second part is to calculate the user ratings matrix on the items of the goods according to the similarity matrix; the third part is the item similarity matrix W and the users of the item score matrix multiplication to obtain the recommendation results. Traditional Item-Based collaborative filtering recommendation algorithm based on item is the stage that affects the performance of the algorithm. If the number of users is n, the number of commodity items is m, the time complexity of finding all the items in the n project is O(m^2), the total search space is n users, so the time complexity of computing similarity is O(m^2n). So when calculating the similarity matrix of items, it is independent of the similarity between the calculated and the other pair of items to a project, so it is possible to calculate the similarity matrix.

2. A new algorithm of Item-based based on MapReduce

The new algorithm is mainly divided into two steps; the first step is the MapReduce implementation of K-Means algorithm based on clustering of users. The second step is to achieve the parallel recommendation algorithm of Item-based on MapReduce, the product of user clustering recommendation.

2.1 The new algorithm K-Means based on MapReduce

The basic idea of the traditional K-means clustering algorithm: from M data objects in arbitrary choice of K objects as the initial cluster centers; for the rest of the other objects, according to their distance and the cluster centers, respectively, they allocated to its most similar clustering; then calculate each received a new clustering algorithm clustering center; keep repeating the process until no changes in a core. In the k-means algorithm to calculate the distance between data objects and cluster centers is the most time-consuming operation. The data object and K cluster center distance comparison at the same time, data from other objects can also be compared with the K distance of the center of cluster, so the operation can be parallelized[8]

Based on MapReduce parallel implementation of K-means algorithm can improve the speed of the clustering algorithm, is divided into three steps: the first step: the map function, for every point calculation recently the center distance and the corresponding to the nearest cluster center. The second step: Combine function, just completed the Map machine on the machine are completed with the same point of the cluster point of summation, reduce the amount of communication and computation of Reduce operation. This step is the key to the use of Combine function on the machine on the first of the same cluster merge, reduced to the Reduce function of the transfer and the amount of computation. The third step: the Reduce function, the intermediate data of each cluster center will be formed and the new cluster center can be obtained. Each iteration is repeated on the three step.

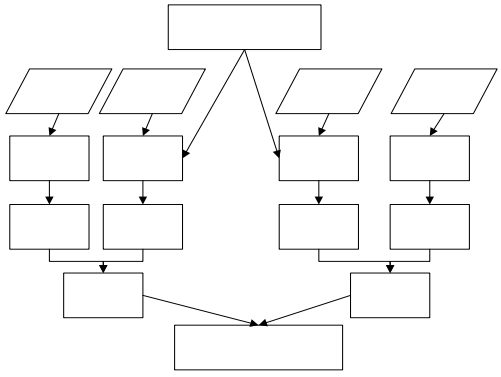

Figure 2 Parallel Flow Chart of K-means Algorithm based on MapReduce

2.2 the collaborative filtering algorithm based on MapReduce for parallel implementation of Item-based

Based on the similarity calculation formula mentioned above (1), this paper presents a collaborative filtering recommendation algorithm based on MapReduce.

Algorithm 1 The collaborative filtering recommendation algorithm based on MapReduce

INPUT: User information file, Item information file, Intended user

OUTPUT: Intended user recommended list

The process is as follows:

Step1: Transforming the user vector into an item vector;

Step2: Parallel calculation of the similarity between items; the calculation of the similarity between items according to the formula (2) to calculate;

Step3: Similarity matrix of parallel computing objects;

Step4: Parallel computing user rating matrix; in the calculation of the user's scoring matrix, if the user is not on the items too much, then the default score is 1;

Step5 :The results obtained by the multiplication of the similarity matrix of parallel computing objects and the user's score matrix are recommended.

Experimental result analysis

1. experimental environment

The simulation experiment using VMware_Workstation_10.0.3, virtualization software to virtual Hadoop cloud platform. Eight virtual machines are installed on the virtual Hadoop cloud platform, and a Hadoop cluster environment is built on these eight virtual machines. One of the virtual machine as a good JobTracker node NameNode, the other seven virtual machines deployed TaskTracker and DataNode. These machines are in the same local area network.

The experiment uses eight sets of virtual machine hardware configuration and software configuration as shown in table 1:

Table 1 Hadoop Cluster Configuration

OS	Centos6.4
JDK Version	1.6.0
Hadoop	1.1.2
HardWare	2GRAM 100G Hard Disk

2. Experiment and analysis

Based on MapReduce parallel implementation of Item-based collaborative filtering algorithm in parallel mode expansion rate performance comparison test, select the size of the data set, respectively, in the efficiency of 1-8 nodes running. The experimental results are shown below:

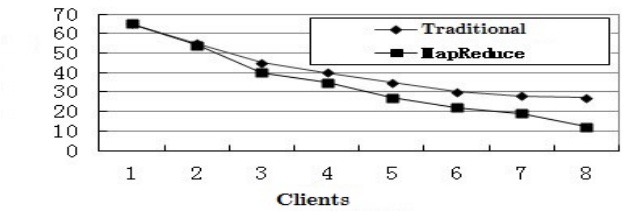

Figure3 Performance Test Chart

Figure 3 is based on MapReduce parallel implementation of item based collaborative filtering algorithm can test chart, the X axis is the number of clients, the y-axis is the response time of the system. The experimental results show that based on MapReduce parallel implementation of item based collaborative filtering algorithm performance compared to the traditional recommendation algorithm is significantly improved.

Conclusion

In this paper, a new algorithm of collaborative filtering algorithm based on MapReduce is proposed. The experiment results show that the new algorithm has high efficiency and can achieve high performance at a low cost. But in this paper, the user clustering is completed on the basis of the user with a small number of attributes, for high dimensional attributes of the user groups, but also to do further research. In addition to the new algorithm in this paper has been put forward, we will continue to improve the experimental method, and constantly improve the accuracy of the recommendation algorithm.

References

[1] Chen ru ming ,Challenges, values and coping strategies in the era of big data[J].Mobile Communications. 2012(17) : 14-15.
[2] Sun Lingfang,Zhang Jing.Electronic recommendation mechanism based on RFM model and collaborative filtering[J]. Journal of Jiangsu University of Science and Technology(Natural Science Edition). 2010,24(3):285-289.
[3] LI Gai,PAN Rong.et Collaborative filtering algorithm parallelize research based on large data sets a[J]. Computer Engineering and Design, 2012,33(6):2437-2441.
[4] LI Wen hai;XU Shuren;Design and implementation of recommendation system for E-commerce on Hadoop[J]. Computer Engineering and Design, 2014(35):131-136.
[5] SUN Tianhao, LI Anneng et.Research on Distributed Collaborative Filtering Recommendation Algorithm Based on Hadoop[J]. Computer Engineering and Applications, 2014,51(15):124:128
[6] Xie Xuelian,Li Lanyou.Research on Parallel K-means Algorithm Based on Clound Computing Platform[J]. Computer Measurement & Control, 2014,22(5):1510-1512.
[7] Yan Cun,Ji Genlin.DesignandImplementationof Item-BasedParallel Collaborative Filtering Algorithm[J].JOURNAL OF NANJING NORMAL UNIVERSITY(Natural ScienceEdition), 2014,37(1):71-75.
[8] WAGN Fei,Qin Xiaolin.Algorithm for k-means Based on Data Stream in Cloud Computing[J]. Computer Science, 2015,42(11):235:239.