Research and Application of Collaborative Filtering Algorithm Based on Hadoop

Ailing Duan[1], Zhixia Xu[2] and Qiongbo Duan[3]
[1]School of Information Science and Engineering Henan University of Technology, Zhengzhou China
[2]Department of water Resources China Institute of water Resources and Hydropower Research
[3]Engineering College of the Armed Police Force Xi'an, Shanxi, P.R. China

***Abstract*—In order to solve the user rating data about data sparseness and traditional similarity calculation method because of its disadvantages of strict match object attributes, combined with the project classification and cloud computing platform is put forward an improved collaborative filtering recommendation algorithm. The experimental results show that the algorithm not only effectively solve the data sparse and the insufficiency of traditional similarity method, but also improved the user's interests and the accuracy of the nearest neighbor search; At the same time, the algorithm only need to calculate where new users or categories, greatly enhance the scalability of the system.**

Keywords-data is sparse; cloud computing; collaborative filtering recommendation algorithm

I. INTRODUCTION

Current predictions recommendation algorithm is filtering recommendation algorithm based on content, collaborative filtering recommendation algorithm, based on demographic recommendation algorithm, based on knowledge of the recommendation algorithm and hybrid recommendation algorithm, which together filtering algorithm is one of the most used one of successful recommendation algorithm[1].

Hadoop is based on distributed storage and parallel computation of cloud computing platform, use of low-cost PC equipment of large cluster, the next generation of high-performance huge amounts of data distributed computing platform, belongs to the completely open source architecture. However due to the large e-commerce site huge and the number of users and products continue to increase, at the same time the user to score a few goods, usually under 1% (5 ~ 9), resulting in data sparse user item rating matrix, seriously affect the quality of the recommendation system, coupled with the inherent cold start and scalability problems of the traditional algorithm. So the method of solving the problem of sparsity, cold start and extension is also appeared. Parallel improved collaborative filtering algorithm is proposed in this paper to solve the problem of traditional algorithms difficult to extend[2], for the collaborative filtering recommendation under the huge amounts of data provides a solution, has a certain reference significance.

II. HADOOP PLATFORM OPERATING MECHANISM

Hadoop core is composed of three parts: HDFS, MapReduce and HBase. HDFS (Hadoop distributed file system)is a run on the cheap hardware Distributed File System, give full consideration to the File System performance, scalability, reliability and availability, and has the function of fault tolerance and automatic recovery. On the data access, the application must be streaming access HDFS data set on it.[3-4]

MapReduce is a parallel and distributed data processing framework in the general computer cluster. the framework of the core idea is: the task is divided into Map and Reduce phase. In the Map phase, each Mapper computing nodes to accept a certain number of data blocks, and then according to the custom Map function, produce < key, value >, and then call the custom Map function for processing and generate the intermediate results in local. In the Reduce phase, remote read the Map phase of intermediate results, call the custom of the Reduce function for processing, and storage to the HDFS will eventually results.

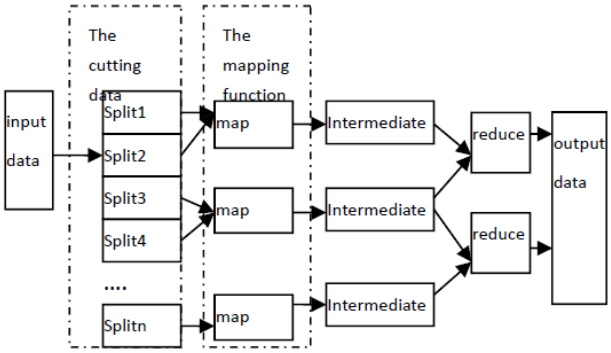

FIGURE I. MAPREDUCE DISTRIBUTED WORKING PRINCIPLE

III. COLLABORATIVE FILTERING ALGORITHM BASED ON THE USER

Collaborative filtering algorithm based on user is the basic idea is through user behavior records, access to information can reflect the user's interests. Then analyze the data and generate a reflects the relationship between users and products score matrix. According to the score matrix, calculate the similarity between users, so as to find similar users. For each user, you can use the similar user interest to predict its interested in commodities, so as to realize recommendation. There are mainly three stages of the algorithm step.

A. The User Information Collection

The user's evaluation into $m \times n$ matrix; Building user - project evaluation matrix r (m, n) to represent the user rating of the project[5], which m number, for the user number n for the

project. Each line represents the user rating of the project, and each column represents different users score for the same project, the default value is 0 users never score of these projects.

$$\begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} \\ r_{21} & r_{22} & r_{23} & r_{24} \\ r_{31} & r_{32} & r_{33} & r_{34} \\ r_{41} & r_{42} & r_{43} & r_{44} \end{bmatrix} \quad (1)$$

B. To Calculate Similarity between Users

All the user preference for some items as a vector to calculate the similarity between the items. Commonly used similarity method with cosine similarity, modified cosine similarity, Pearson correlation, and other methods. Each of the formula has certain difference. In practice, mainly to see how much the amount of data. Large amount of data, then calculated the similarity of the closer.

1) Cosine similarity: In the user - project matrix r, the user is expressed as a one-dimensional vector, the similarity between the user can use the cosine values of the two vectors. Cosine values of [0, 1], the cosine value, the greater the similarity is higher.

$$w_{i,j} = \cos(\vec{i}, \vec{j}) = \frac{\vec{i} \bullet \vec{j}}{\square \vec{i} \square \bullet \square \vec{j} \square} \quad (2)$$

$\vec{i}$ $\vec{j}$ representative user score vector.

hypothesis $\vec{i} = \{x_1, y_1\}$ $\vec{j} = \{x_2, y_2\}$ (3)

Cosine similarity of users i and j.

$$w_{i,j} = \cos(\vec{i}, \vec{j}) = \frac{\vec{i} \bullet \vec{j}}{\square \vec{i} \square \bullet \square \vec{j} \square} = \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2}\sqrt{x_2^2 + y_2^2}} \quad (4)$$

2) Modified cosine correlation: Cosine correlation does not take into account the differences of different user ratings, modified cosine correlation can be minus the average user ratings to improve the defect.

C. Nearest Neighbor and Forecast

Will the rest of the user according to similarity from big to small sort, and find out the largest use of k similarity as recommended the nearest neighbor; With k nearest neighbor similarity and nearest neighbor score value value to predict the user of the program of the project value[7-8].

User u prediction score on project i Pu i, the simplest method is to use $u_{neighbour}$ to the project i the mean score of k nearest neighbors (formula 3-5), but this way predicted results is not good enough, did not consider similarities with your neighbors. Now commonly used is to consider the user similarity forecast method (formula 3-6), compared with considering the difference of user ratings.

$$p_{u,i} = \frac{\sum_{u_k \in u_{neighber}} r_{u_k, i}}{k} \quad (5)$$

$$p_{u,i} = \frac{\sum_{v \in u} w_{u,v} r_{vi}}{\sum_{v \in u} |w_{u,v}|} \quad (6)$$

According to characteristics of algorithm, three problems in traditional algorithm. Data sparseness:

1) On large e-commerce sites, each user comments by the number of not more than 1% of the total number of projects. In this case, the score matrix is extremely sparse, there will be two have similar users by common grade project little and similarity to zero, this situation is called neighbor transmission loss. Collaborative filtering algorithm mainly rely on users to recommend resources score, score a lack of will affect the similarity between the user of the project, causing the wrong recommended. In this paper, the related improvement data sparseness.

2) The cold start problem is also a classic problem of collaborative filtering algorithm, divided into new users, the new project. New user problem refers to a system for users joined system without any project of score can't calculation for its neighbor, also cannot be recommended for its projects. New project problem refers to just add new project due to the lack of scoring can't recommend to the user.

3) Scalability problems: similarity calculation and nearest neighbor search complexity, the highest and most time-consuming. As can be seen from the Figure 2, the server side under the recommended all computing tasks. The most common solution is to improve the performance of the server, or distributed clustering is used to increase the processing capacity.

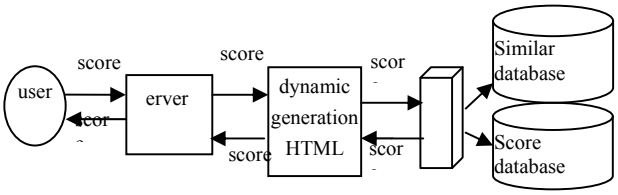

FIGURE II. SERVER WORKING PRINCIPLE

IV. HADOOP PLATFORM PARALLEL IMPROVED COLLABORATIVE FILTERING ALGORITHM

Only for data sparseness improvement is not enough, the collaborative algorithm also faced the extension. To solve the problem below. Using the Hadoop has a strong processing capacity, and can be parallelized code only needs to implement Map and Reduce class. Begin from collaborative filtering algorithm, analysis algorithm process, the traditional collaborative filtering algorithm is decomposed into data, nearest neighbor, score[10], attribute weights of forecasting the four parallel Job tasks, Job order between the two, but each Job is independent of the task. Over a period of time, can be thought of as the user's interest is the same, users increased by a small amount of scoring record in the short term will not

affect the calculation of similar adjacent. Therefore data combing, similarity computation, the weights of attributes to predict the three jobs do not need to perform every time.

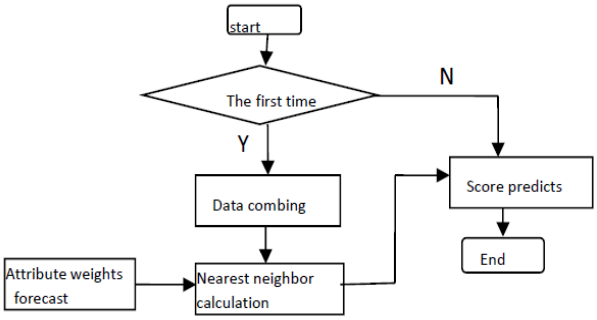

FIGURE III. COLLABORATIVE FILTERING ALGORITHM FLOW CHART

As can be seen from the fig3, parallel collaborative filtering can be divided into two parts of the online and offline. Combing offline part includes data, similarity calculation, attribute weights of prediction of three parts. Combing the main data processing to comb of huge amounts of data, implement classified by user id; Nearest neighbor is computing similarity between users, nearest neighbor; Property rights value forecast was primarily based on user ratings records, calculating the user on a particular attribute weights, based on the attribute value prediction is not score score value of the project. Online part is score predicts, predict users never score score value of the project.

A. Data to Comb

Under the huge amounts of data, don't put all the data in the memory. No correlation between the data, the data can be assigned to different execution node, in turn, improve the execution efficiency. According to hadoop programming process, the comb into the Map data, Combine, Reduce three stages.Map phase to accept the file block, according to the line read each record, to deal with each record, user id, item id, the score values. After processing, the user id as the key, project id and score value as the value, the output < key, value >.Combine phase is in order to alleviate the Map inter-node communication burden, the large amounts of data generated in the same node number data by a hyphen is connected with a common key, to Reduce function after partial results are obtained. Is the same user id as the key, project id and score value as the value, the output < key, value>. Reduce phase will be treated as multiple Map data integration processing, each all of the user's score values. Is the same user id as the key, the project id and score value as the value.

B. Similarity Algorithm is Described

On Hadoop platform, the similarity calculation allocated more child nodes, use the heap sort of nearest neighbor. Nearest neighbor can use original collaborative filtering algorithm calculates the cosine value of between users or Pearson, value, or similarity and attribute weights based on user prediction algorithm to calculate the nearest neighbor, and need to calculate in advance each user preference for attribute weights. Calculate preference weight can also through the Hadoop platform parallel implementation. TopK nearest neighbor to determine the minimum heap by building a capacity of k, by adjusting the heap of maximum k nearest neighbor. Particular way, the first design three classes to implement the neighboring calculation. Comb design SimilarityDriver class is a data entry, first of all, the Job SimilarityDriver class is initialized, set up the Map, Reduce the concrete implementation of the operation. Results can be set by adding the input file directory function, the output directory. Second, and Reduce design SimilarityMapper class make Map node is communication through the network to realize the interaction, in order not to cause the network transmission delay is too large, as far as possible, Reduce the transmission of data. So in the design of function, the computing process to a Map function will be as much as possible. In SimilarityMapper class, not only to calculate a particular user similarity with other users, also calculate topK nearest neighbor. At last, through design SimilarityReducer class Map generated k nearest neighbor, and write the final result to HDFS, calls for score predicts. Specific SimilarityMapper class code description is as follows:

Input: < key, value >, including key said starting line offset, value represent user ID value, k neighbor number.

Output: < key, value > , including key said for ID, said the value the user's nearest neighbor, including neighbors ID, similarity and other information. Steps are as follows:

Step1: Use data combing the data to initialize the score matrix, and create a new capacity of K minimum heap;

Step2: Retrieve the user ID value from the value of the uid; Calculate the uid and the similarity between users;

Step3: Put x user in a minimum heap, and according to the similarity value adjustment heap;

Step4: Cycle through the uid with other users are similarity calculation;

Step5: Uid as the key, the inside of the minimum heap neighbors as a value;

Step6: Output <key,value>

END

SimilarityReducer class code description is as follows:

Input= output = < key, value >, of which the key user id, Value according to the user the corresponding k nearest neighbor.

Step1: define a variable result;

Step2: Read the value from the iterator values, and add the result to the result, until completion of the iterator read;

Step3: The result as a value, and through the context output;

END

C. Score Prediction Algorithm

A grade prediction mainly through similar adjacent values to forecast of users never rated items, score predicts the nearest neighbors in the nearest Job has been calculated. Nearest

neighbor specified in the score predicts Job initialization phase calculation Job output paths, in the Map phase computing user u have not rated items j to score, in the Reduce phase output prediction score values. On request, can give more child nodes will score predicts request load, so it can achieve good response to the user's request. Specific design two ForecastMapper and ForecastReducer. Through design ForecastMapper class implements the map process, according to the reading of records, analysis the user id and project id. After get the user id, read the user from a second Job output path u set corresponding to the nearest neighbor, and also need to read from the first Job score value of the nearest neighbor for project j. Secondly by designing ForecastReducer class to predict the score value output to the HDFS file. System according to the output rating value for project recommend users.

D. The Weights of Attributes

Combing the original data classified according to the user id, assign each user's scoring record child nodes perform the Map operations. In the operation of the Map to weight training of users, in Reduce operating output each user attributes weights. As before design two BWAPredictMapper and BWAPredictReducer, through BWAPredictMapper class implements the map process. Record user ratings classified by Id, and the average user assigned to compute nodes for user weight training. Finally through BWAPredictReducer class just weights of the user information integration, and then output to the HDFS.

V. THE EXPERIMENTAL RESULTS

Using the Hadoop cluster system consisting of six computer interconnection experiment. Experiments will be one of the computer as a name node (the NameNode), it is the core of the HDFS, don't participate in operation; The remaining 5 computer DateNode and TaskTracker service node for the data. The experimental hardware environment for Intel Centrino2.5 Ghz CPU and 1 gb of memory, configuration based on Hadoop2.3.0 version of the cluster system. Using graphs programming model and the realization of JAVA language programming algorithm.

In order to verify the algorithm's extensibility, adopt the method of a closed TaskTracker, one by one to reduce computing nodes is to change the processing power of clusters.

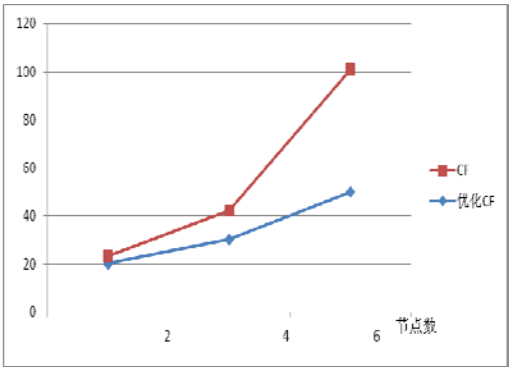

FIGURE IV. COLLABORATIVE FILTERING ALGORITHM COMPARISON CHART

This experiment data size: 1.88 M, 23.4 M, 254 M. The results show that based on Hadoop collaborative filtering algorithm has good expansibility, when processing node in the cluster number increase, the shorter the time, when the data set, the greater the increase significantly.

VI. CONCLUSION

Aiming at the shortcomings of traditional collaborative filtering algorithm, put forward specific measures for the improvement of traditional collaborative filtering algorithm. Articles with big data simulation experiments, the experimental results show that the improved algorithm on the recommended recommend efficiency and precision has obvious advantages. With the development of personalized recommendation, at the request of real-time and complexity of recommendation algorithm will be the focus of the recommendation algorithm research in the future.

REFERENCE

[1] Yan wei. The data mining algorithm based on cloud platform research and implementation [D].Cheng du: University of electronic science and technology, 2013-03-01.

[2] Wang zhenyu,li Guo. An Analysis of the Search Engine User Behavior sBased on Hadoop.COMPUTER ENGINEERING & SCIENCE.2011.33(4).

[3] Huyu, Feng jun. Distributed Search Engine Using Hadoop.COMPUTER SYSTEM & Application.2010.19(7).

[4] Hang bin,Xu shuren. Design and implementation of MapReduce-based data mining platform.[J]. COMPUTER ENGINEERING AND DESIGN.2013.34(2)495-501.

[5] Taojun,Zhang ning, Collaborative Filtering Algorithm Based on Interest-Class.[J]. The computer system application.2011.20(5)55-59.

[6] duo xuesong,zhangjing etc. A Mass Data Management System based on the Hadoop. COMPUTER INFORMATION.2010.26(5-1)

[7] Hadoop Wiki, http://en.wikipedia.org/wiki/Hadoop

[8] Tom White. Hadoop: The Definitive Guide. O`RELLY Press.

[9] MapReduce: Simplified Data Processing on Large Clusters. Google Inc.

[10] Wang qian,Wangjunpo. An improved collaborative filtering recommendation algorithm.[J]. computer science.2010,37(6):226−227