Research on Improved Collaborative Filtering Recommendation Algorithm on MapReduce

Jie DONG[1,a], Yun QIN[1,2], Xue Yang SUN[1]and Li Ming DU[1]

[1]*Information and Control Engineering College, Shenyang Jianzhu university, Shenyang, 110168, China*

[2]*Liaohe Oilfield Company SAGD Project Department*

Abstract. Information overload is one of the most serious problems in big data environment, recommendation systems is a way to effectively mitigate the problem. In order to make use of rich user feedback and social networks information and to further improve the performance of the recommendation system ,This thesis makes a improvement on the user-based collaborative filtering algorithm by normalization method, Meanwhile the algorithm could be run on the MapReduce in the Hadoop platform. The experimental results show that the algorithm on Hadoop platform can effectively improve the accuracy of the data to recommend and computational efficiency, so as to improve the satisfaction of users.

1. Introduction

The rapid development of the Internet have brought about great changes in the world, which cause information overload[1-3]. Information overload has two sides .For information users, searching for what they need from the vast amounts of information accurately is becoming more difficult .For information manufacturers, making information they produce conspicuous in the vast amounts of information is also a problem need to be solved urgently. Recommendation system is based on the analysis of used behavior to meet users' need and interest information for the user in initiative.

In 2003 Deng Ailin put forward a collaborative filtering recommendation algorithm based on project score predicts [4] in order to solve the sharply fallen recommendation quality rating caused by extremely thin data .In 2004, Gao Fengrong adopt the method of clustering and classification respectively by divided sparse score matrix, reducing the scope of neighbor and the need to predict the number of resources, improving the efficiency and scalability of the collaborative filtering recommendation algorithm [5]

With the growing amounts of information, Research focusing on the recommendation system has changed from stand-alone mode to server cluster [6-8]

In 2011, Jing Jiang and others further divided collaborative filtering recommendation algorithm based on the project into three stages of the main computing [7].After the segmentation of each Map Reduce phase can be run in parallel in each node of the cluster. Meanwhile they put forward the data partitioning strategy based on Hadoop platform to maximize the local data storage, reducing the communication overhead between system and improving the recommendation efficiency of the recommendation system. In 2012, Juan Yang and others further used mongo as auxiliary storage database [8].

At present, collaborative filtering recommendation methods are mainly based on matrix decomposition [9][10]. However, the current method in scalability performance is poor, and unable to adapt to the massive data processing.

In this paper, aiming to the recommendation system's error range,. We analyzed deeply the traditional collaborative filtering recommendation algorithm based on user, run the improved the algorithmon on MapReduce in the platform of Hadoop ,which has made the effect more accuracy

2. The introduction of collaborative filtering recommendation algorithm

Through the ten year research, a variety of recommendation algorithm has been proposed and collaborative filtering recommendation algorithm is widely used.

Collaborative Filtering Recommendation algorithm usually uses the nearest neighbours technology and utilize the history of users' preference information to calculate the similarity between users' interest and use the nearest neighbours of the target users by items evaluation value to predict the target user preferences for certain products to recommend for target users. The largest advantage of collaborative filtering algorithm have no special requirements on recommend objects And it also can handle unstructured complex objects such as

[a] Corresponding author: dong_jie_2002@163.com

music and movies. At present, collaborative filtering has been used in e-commerce recommender systems by based on other users' evaluation to recommend for target users. Collaborative filtering recommendation algorithm are usually divided into two types: Item-based and user-based. The following will be introduced to these recommendation algorithm respectively.

2.1 User-based collaborative filtering recommendation algorithm

Collaborative Filtering Recommendation algorithm usually uses the nearest neighbors technology and utilize the history of users' preference information to calculate the similarity between users' interest and use the nearest neighbors of the target users by items evaluation value to predict the target user preferences for certain products to recommend for target users. Specific process is as follows:

Step1: Collection and storage, sorting out the history of user's behavior data such as the user's existing purchase records, browse list page, the collection behavior, attention, project ratings, etc. These data is the base for calculating the similarity of users.

Step2: Searching for collection of users similar to target users, the key to this step is calculating the similarity between two users.

In order to calculate user similarity, Cosine similarity calculation is often used to calculate the degree of similarity between two users. Assuming $\vec{i}$ and $\vec{j}$ represent n dimension vector of user *i* and *j* respectively, the similarity between the user *i* and *j* is defined as follows:

$$sim(i,j) = \cos(\vec{i},\vec{j}) = \frac{\vec{i}\cdot\vec{j}}{\|\vec{i}\|*\|\vec{j}\|} \tag{1}$$

In practice, the cosine similarity did not take into account user ratings subjective preference. cosine similarity is modified to solve the above problem. It subtracts the user average score for the item so as to reduce difference of the subjectivity. Assuming $I_{i,j}$ is scores set that user *i* and *j* give the items together, I_i and I_j are scores sets of users *i* and *j* respectively, $R_{i,c}$ or $R_{j,c}$ respectively stand for score of user *i* or user *j* on the item c, $\overline{R}_i$ and $\overline{R}_j$ are users *i* and *j* average' scores respectively, the user similarity between user *i* and user *j* is as follows:

$$sim(i,j) = \frac{\sum_{c\in I_{i,j}} (R_{i,c}-\overline{R}_i)(R_{j,c}-\overline{R}_j)}{\sqrt{\sum_{c\in I_i}(R_{i,c}-\overline{R}_i)^2}\sqrt{\sum_{c\in I_j}(R_{j,c}-\overline{R}_j)^2}} \tag{2}$$

The accuracy of the similarity calculation directly decides the recommendation algorithm quality.

Step3: By means of similarity calculation results, the most nearest neighbors of target user could be made of the top N users which similarity value are very similar to the target user similarity degree, at this time the value N can be set according to the need to manually. After that ,we can obtain recommend results for the target user .

Supposing N_u represent a nearest neighbors set to target user , $P_{u,c}$ represent a score of user u on the item c, calculating method as shown followed:

$$P_{u,c} = \overline{R}_u + \frac{\sum_{v\in N_u} sim(u,v)*(R_{v,c}-\overline{R}_v)}{\sum_{v\in N_u}|sim(u,v)|} \tag{3}$$

2.2 Item-based collaborative filtering recommendation algorithm

Item-based collaborative filtering recommendation algorithm based on the similarity between items, finding the nearest neighbors set of items, according to the user preference information to predict item ratings, eventually object could be feedback to the user by those similar items which users like , but Item-based collaborative filtering recommendation algorithm does not use of the attributes contents of the item to calculate the similarity between items, but to calculate the similarity through the analysis of user behavior records.

3. Distributed Computing Framework Based on MapReduce

Under the circumstance of the big data, Hadoop is one of the basic infrastructure which make full use of the cluster ability of distributed computing. MapReduce is a programming framework which handle a large number of data based on the hadoop platform, and it is also a kind of application in large-scale parallel processing of information, a new distributed computing model. No matter how many hosts on Hadoop platform cluster, MapReduce technology can finish the tasks such as web searching, advertising pushing and rapid processing even for TB, PB, and even EB bytes of data on thousands of cluster nodes at the same time.

The core idea of MapReduce algorithm is dividing and then rule it, namely the operation of large-scale data sets, firstly distributed to various points node to complete together ,which is under the management of one master node, then getting an intermediate result. After that integrating the middle results of each node, and obtaining the final result . The entire MapReduce framework is composed of the Map and Reduce functions, run map stage first, then perform reduce stage. The Map is responsible for dividing a task into many subtasks and divided large-scale data into several smaller blocks, then transfer to a large number of cluster nodes. The output of map stages is the input of the Reduce phase. Reducing is responsible for integration and merge and then the child tasks processing results.

4. The Improvement to User-based Collaborative Filtering Recommendation Algorithm

4.1 The improvement of computing user similarity

The formula (1)-(3) show that the user similarity computing depends on the score of items given by users. In fact, some users are very casual, and all scores are higher. And some users are very strict with it, even if the quality of the goods is well; however the scores are also not high. The more the difference of evaluation standard is, the worse the accuracy of recommendation.

Although modified cosine similarity take into account above problem, but it still slightly rough. In this paper, the subjective normalization algorithm is introduced. By using the initial users-item evaluation scores, normalized scores of user-item are calculated in order to eliminate the error that caused by user different grading standard. The specific implementation process is shown below.

Step1: according to the initial scores, calculate each user's average score r_u, calculation method is as follows:

$$\overline{r_u} = \frac{\sum_{i=1}^{n} r_{u,i}}{n} \tag{4}$$

In the formula, $r_{u,i}$ is the score given by the user u to the item i.

Step2: set user ratings interval by calculating averages of the minimum and maximum of all user scores as both ends of the interval, and calculation method is as follows:

$$[\frac{\sum_{j=1}^{m}\min(r_{u_j,i_1}, r_{u_j,i_2},.....,r_{u_m,i_n})}{m}, \frac{\sum_{j=1}^{m}\max(r_{u_j,i_1}, r_{u_j,i_2},.....,r_{u_m,i_n})}{m}] \tag{5}$$

$$b_{i1} = \min(r_{ui,j1}, r_{ui,j2}.....,r_{ui,jn}) \tag{6}$$

$$b_{i2} = \max(r_{ui,j1}, r_{ui,j2}.....,r_{ui,jn}) \tag{7}$$

$$k = \frac{\sum_{i=1}^{m} b_{i1} + \sum_{i=1}^{m} b_{i2}}{2m} \tag{8}$$

In the above formula, b_{i1} and b_{i2} stand for the minimum and maximum values by calculating each user the boundary value, k is average for all users interval.

Step3: recalculate score $r'_{u,i}$ that the user u give the item I in order to obtain a new user -item evaluation.

$$r'_{u,i} = \frac{k * r_{u,i}}{\overline{r_u}} \tag{9}$$

In formula, the normalized coefficient is $\frac{k}{\overline{r_u}}$.

By the formula (9), via the derivation transformation, the average score of each user has become interval average k, namely: $\overline{r'_{u1}} = \overline{r'_{u2}} = ... = \overline{r'_{um}} = k$

Using this normalization processing on the user's score, it can unify the standard of the evaluation. And it can effectively reduce impact which the different user rating scales on the similarity computations.

4.2 Improved recommendation algorithm for scoring prediction algorithm

When we calculate the user similarity by the traditional collaborative filtering algorithm, it can calculate on the item set that the user has scored together. If two users give the more same type of items, however the grade is less for the same item. They are similar to each other in theory, but according to the traditional similarity calculation formula, it will leads to the dissimilarity results between two users. Therefore, on the basis of the subjective normalization method, prediction score calculation method is further introduced. When we calculate the similarity between two users, we can increase user's common grading evaluation scores to reduce data sparse and improve the quality of recommendation.

The concrete implementation process of improved filtering recommendation algorithm is shown below.

Step1, according to user-item evaluation matrix A (m, n), we are able to calculate union sets of item evaluation among each users ,calculation method as follows:

$$I_{i\cup j} = I_i \cup I_j \tag{10}$$

Among the formula: I_i and I_j are sets that users i and j have evaluated scores respectively, $I_{i\cup j}$ is a union set between the user *i* and *j*.

Step2 using subjective normalization algorithm to construct normalized score matrix, calculation method is as follows:

$$A(m,n) \to A'(m,n)$$

Step3 according item set $I_{i\cup j}$ of users *i* and *j* after users-item evaluation matrix normalized calculation $A'(m,n)$ to calculate similarity, the calculation method of the improved formula(3) as follows.

$$sim(u_i,u_j) = \frac{\sum_{c\in I}(\mathrm{r}'_{u_i,c}-k)(\mathrm{r}'_{u_j,c}-k)}{\sqrt{\sum_{c\in I}(\mathrm{r}'_{u_i,c}-k)^2\sum_{c\in I}(\mathrm{r}'_{u_j,c}-k)^2}} \tag{11}$$

Among the formula: *k* is the normalization factor, *I* represents union set that the user *i* and user *j* have evaluated items commonly by the formula (3)

Step4 assumes the user u_a is the target user, c is for forecasting project, improve user's similarity calculation formula (4) to calculate the forecast scores of user u_a on item *c*, calculation method is as follows:

$$P_{u_a,c} = k + \frac{\sum_{u\in U} sim(u_a,u)(r_{u,c}-k)}{\sum_{u\in U}|sim(u_a,u)|} \tag{12}$$

5. The realization of the improved algorithm based on MapReduce

In the Collaborative filtering recommendation algorithm based on the user, first of all, according to the user-item evaluation matrix the degree of similarity among the projects are calculated. Then on the basis of similarity, calculating similarity between the user and the target

users，predicting scores which target users do not have the evaluation score of the item, and obtaining the TOP-N value of highest score. Obviously these two steps is a serial process of order.

In the process of user similarity calculation, there is almost no coupling between them, so it can be parallel computing, which could assign a MapReduce job to implement. In the prediction of target users evaluation score ,it also can use MapReduce job parallel computing forecast rating of each item.

Thus, the essence of collaborative filtering recommendation algorithm based on the user for distributed computing is to calculate the similarity between users with a MapReduce task. And its result is another input of MapReduce task to calculate the target users forecast rating of the item. And the two tasks are executed serially.

Subjective score normalization algorithm is mainly responsible for the user-item rating matrix is converted to a unified coordinate system, the parallelization of subjective score normalization algorithm is divided into three steps. The first step is according to the user to solve the user-item rating matrix. The second step is according to the user-item rating matrix and solving the maximum ，minimum and average score. The third step is based on the user-item rating matrix and maximum, minimum and the average score of solving normalized user-item rating matrix.

In the Improved User-based Collaborative Filtering Recommendation Algorithm, Each of three processes, which are user evaluation matrix normalized, users prediction score, calculate similarity .They can be realized in parallelization, these three parallelization process combined into a serialization process. Each intensive computing process segment with parallel ideas to speed up the whole operation process, and make the algorithm tolerated normal time to complete the calculation. Through the above steps, the parallelization of the recommendation algorithm has been realized, which reduce the running time of the recommendation algorithm and improve the efficiency of the recommendation system.

6. Test results

The experiment is needed so as to prove the effectiveness of the improved collaborative filtering recommendation algorithm based on Mapreduce, In the experiment, using the data of 6000 users to 1000000 scores of 4000 movies information and inputting the user ID and the recommended number, in the end the list of recommendation are displayed. For example to experiment , supposed the users NO.6 and the recommended list number is 20, the experimental results are shown in figure 1, in which the vertical axis shows recommended time used, the horizontal axis shows node number

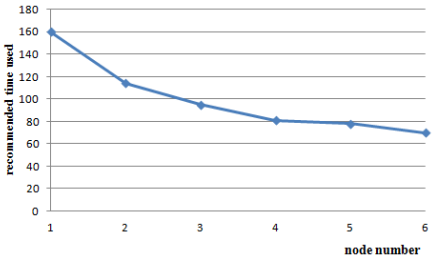

Fig1. Recommended performance comparison

Experimental results show that as the number of nodes increases unceasingly, Using the improved algorithm, it still get better recommendation results. recommended time used decrease, which can illustrate the performance of the system are on the increase and increase accuracy of the recommendation and the efficiency of recommendation, reduce the impact of the sparse data.

7. Conclusion

Aiming to the bottlenecks problem of traditional recommendation algorithm operations on a single machine performance, the improved collaborative filtering recommendation algorithm on MapReduce framework in the Hadoop distributed platform are put forward, Experimental results show that it increases recommending performance, which the computing tasks can be evenly distributed to each computing nodes in distributed parallel processing, and has the good expansibility.

Acknowledgement

In this paper, the research was sponsored by the Nature Science Foundation of Liaoning Province (Project No. 2014020068)

References

1. Yuanzhuo Wang, Xiaolong Jin. Journal of Computer, **36**,4(2013)
2. Jianguo Liu, Tao Zhou, Binghong Wang. Progress in natural science, **19**,7(2009)
3. Shahabi C,Chen Yishin. Distributed and Parallel Databases,**14**,9(2003)
4. Ailin Deng, Yangyong Zhu, Bole Shi. Journal of software. **14**,8(2003)
5. Fengrong Gao, Xiaoyong Du, Shan Wang. Microelectronics and computer. **21**,6 (2004)
6. Zhiyou Zhang. Laboratory research and exploration. **25**,3(2006)
7. Jing Jiang, Jie Lu, Guangquan Zhang, Guodong Long. . Services (SERVICES), 2011 IEEE World Congress . 8(2011)
8. Juan Yang, Han Du, Bin Wu, Xinxin Ge. CCIS, 2012 IEEE 2nd International Conference **1**,5(2012)

9. Ming-Sheng Shang, Linyuan Lü, Wei Zeng, Yi-Cheng Zhang, Tao Zhou. EPL(Europhysics Letters). **88**,3(2009)
10. E Haihong, Song Meina, Li Chuan，et al. Journal of Beijing University of Posts and Telecommunications.**37**,5(2014)
11. Ren Xiuchun, He Yaji. Electronic design engineering, 22, (2014)