Research of User-Based Collaborative Filtering Recommendation Algorithm Based on Hadoop

Y.L. Zhang
School of Computer Science&Engineer
Xi'an University of Technology
P.R.China

M.M. Ma
School of Computer Science&Engineer
Xi'an University of Technology
P.R.China

S.P. Wang
School of Science
Xi'an University of Technology
P.R.China

***Abstract*--Collaborative filtering algorithm is one of the key technologies of the current e-commerce recommendation system, in which the effect of similarity measure directly determines the accuracy of the recommendation system. An improved method of similarity measure and the corresponding collaborative filtering recommendation algorithm by introducing a common user items popularity rating between features and user relevance is proposed in this paper. Furthermore, the implementation of collaborative filtering recommendation system based on hadoop is discussed. The respective evaluation of traditional collaborative filtering recommendation algorithm and improved recommendation algorithm by using MAE show that the improved algorithm enhances the recommendation accuracy in a certain extent. Meanwhile, the overall performance experiments show that collaborative filtering recommendation engines continue to reduce the calculation time with the appropriate increase of the virtual machine node.**

Keywords-collaborative filtering; similarity measure; recommended system; hadoop

I. INTRODUCTION

The work of recommendation systems is to collect and analyze existing behaviour information of the user, in accordance with the relevant recommendation algorithm to generate recommendation of the target user[1].Among many recommender systems, collaborative filtering recommendation is by far the most widely recommended technique[2].The basic idea is that the target users are interested in the favourite goods of neighbours who have common interest with them, which uses a similarity measure algorithm searches neighbours have a common interest with them and generates candidate recommender item by the screening information of neighbour user , and finally recommend it to the target user.

In collaborative filtering recommendation system, how to calculate the similarity score matrix between items is a core part of the user-based collaborative filtering recommendation algorithm. MENG Xian-fu[3] proposed a collaborative filtering recommendation algorithm based on Bayesian theory, which analyze user on item eigenvalues preference using Bayesian theory. Zhao Hongxia[4] put forward the analysis of hybrid collaborative filtering algorithm factor of user and item, which uses the method of factor analysis of data dimensionality reduction, and the use of regression analysis to predict the value to be assessed. Yang Yang[5] proposed a recommendation algorithm based on matrix decomposition and the model of user neighbour . Hao Liyan [6] proposed a sparse filling scoring matrix, calculated on the complete filling matrix similarity and the user's confidence in the similarity factor is introduced to improve the quality of recommendation. However, this algorithm is not suitable for large-scale recommendation systems.

The paper is discussed from the perspective of making full use of user information, considering common rating items and its popularity in the similarity calculation and the introduction of user eigenvalues impact factor of similarity, then the improved similarity calculation method and as a basis for collaborative filtering algorithms are proposed. In this paper, based on hadoop implementation techniques of collaborative filtering algorithms, through the development and test of real systems, indicate recommendation algorithm accuracy and efficiency of the system are improved.

The rest of this paper is organized as follows: Section 2 discusses the traditional user-based collaborative filtering algorithm. Section 3 describes the improved user-based collaborative filtering algorithm. Section 4 presents the implementation of the user-based collaborative filtering recommendation algorithm based on hadoop. Section 5 shows our experiment results and our analysis. Section 6 concludes the paper.

II. ANALYSIS OF TRADITIONAL SIMILARITY CALCULATION METHOD

For the target user u, nearest neighbour recognition is to produce a set of neighbour similarity descending. In multiple collaborative filtering algorithms, cosine similarity, related similarity and modified cosine similarity are used. But those

algorithms have the following two issues in calculating the user similarity.

First, similarity calculations only consider similarity of common user rating items, and ignoring the popularity rating items. Taking book as an example item, although two users have bought “Xinhua Dictionary”, it does not mean the interest of them are same because, the book is just a tool, "Xinhua Dictionary" is the vast majority of Chinese people have bought; However, if two users have bought "Introduction to Data Mining", it can be considered quite similar to their interests. In other words, two users on popular items to take over the same behaviour better able to explain the similarity of their interest, which is taken into account in the calculation of the similarity of the popularity factor of project evaluation.

Secondly, there may be “cold start” problem of the similarity of the results. The so-called "cold start" problem refers to newly registered users do not have any projects been evaluated, and then the similarity between that user and other users is very low, less than the threshold. In this case, the system will not achieve good recommendation to the user effect.

Through the above analysis, the traditional similarity measurement methods cannot effectively measure the similarity between the targets, making the recommendation system cannot find the exact target of the nearest neighbours, leading to recommend quality.

III. IMPROVED USER-BASED COLLABORATIVE FILTERING ALGORITHM

Improved user-based collaborative filtering recommendation is based on the user's score matrices and user demographics property, produce results on the target user's recommendation, which similarity is calculated by a modified cosine similarity algorithm. First, based on resolving item popularity and cold start an improved the similarity calculation method is given, then proposed user-based collaborative filtering algorithm based on the similarity calculation.

A. Improved Similarity Calculation Method

In the traditional recommendation systems, the calculation of similarity does not take into account the popularity of common user rating items , this will lead to more popular when the user merchandise, movies, music produces an almost equal scores, their similarity with tradition similarity measure results would be very high.

First, Adding popularity of common user rating items in similarity calculations, Popularity punishment coefficient is defined as $1/\log(10+\mathrm{N(i)})$,wherein $\mathrm{N(i)}$ is the number of times that i-th item has been evaluated.

Second, for recommendation systems that may appear "cold start" problem, the user's demographics (such as age) add to the similarity relationship eigenvalues calculation,the greater the Age attribute value gap, the greater the gap between the user's interests. So it's calculate equation of age relationship as follows

$$sim_{age}(u,v)=\frac{1}{\sigma\sqrt{2\Pi}}e^{-\frac{(A_V-A_U)^2}{2\sigma^2}} \tag{1}$$

Based on the above ideas, it’s similarity calculation formula as follows

$$sim(u,v)=(1-\lambda)\frac{\sum_{i\in M_{uv}}(R_{u,i}-\bar{R}_u)\times(R_{v,i}-\bar{R}_v)\times\frac{1}{\log(10+N(i))}}{\sqrt{\sum_{i\in M_u}(R_{u,i}-\bar{R}_u)^2}\times\sqrt{\sum_{i\in M_v}(R_{v,i}-\bar{R}_v)^2}}+\lambda\frac{1}{\sigma\sqrt{2\Pi}}e^{-\frac{(A_V-A_U)^2}{2\sigma^2}} \tag{2}$$

B. User-Based Collaborative Filtering Algorithm

In this section, the user-based collaborative filtering algorithm is proposed based on the improved similarity calculation method.

Algorithm input: user-item rating matrix R, item set $I=\{i_1,i_2,\cdots,i_M\}$, user set $U=\{u_1,u_2,\cdots,u_N\}$,recommended set number of elements p, the target user u.

Algorithm output: recommended set I_{rec}.

Algorithm definition:

Calculation of the user u not rating item set $N_k=I-I_k$ （1≤k≤M）, I as item set, I_k is the item set have been rated by user u.

User’s similarity calculations. The results of using (5) computing pair wise similarity between users are stored in user similarity matrix $R_{sim}(N,N)$.Where, M_{uv} indicates user u and v common ratings over a set of items, $R_{u,i}$ indicates user u rating over item i, $\overline{R}_u$ indicates that the user u mean scores for all items evaluated. $\mathrm{N(i)}$ is the number of times that i-th item has been evaluated. A_u indicates the age of user u. λ indicates relationship between eigenvalues in user weights in similarity calculation.

Nearest neighbour set $U_p=\{u_1,u_2,\cdots,u_p\}$ of the target user u is obtained according to the formula $R_{sim}(N,N)$, this makes $u\notin U$ and exists $sim(u,u_1)$ is maximum, $sim(u,u_2)$ follows, and so on.

It calculates and obtains the recommendation set I_p according the nearest p neighbour set U_p of the target user u. Each item i in no-rating item set N_k of user u use (3) to predict user u on item scores [7].

$$P_{u,i}=\overline{R}_u+\frac{\sum_{u_k\in U_p}sim(u,u_k)(R_{u_k,i}-\overline{R}_{u_k})}{\sum_{u_k\in U_p}(|sim(u,u_k)|)} \tag{3}$$

wherein: $sim(u,u_k)$ indicates the comprehensive similarity between user u and his nearest neighbour u_k, $R_{u_k,i}$ indicates user u_k rating over item i. $\overline{R}_u$ and $\overline{R}_{u_k}$ indicate mean scores for items of user u and user u_k respectively.

The elements in N_k for predicting scores are sorted descending order, then taking the first p item consisting of recommendation set recommend the target user u.

IV. User-Based Collaborative Filtering Recommendation Algorithm Based on Hadoop

User-based collaborative filtering recommendation algorithm based on hadoop user parallel idea to take in the calculation process, its main job is to design and implement the Map and Reduce functions. Recommendation algorithm is the most important measure of similarity, pseudo parallelized codes of similarity algorithm shown in Figure 1.Assuming sim(pair, current) indicates modified cosine similarity computation algorithm, age(key, current User) indicates the computation algorithm to calculate age-related degrees.

Input: user datasets,item rating datasets,target user.
Output: similarity.
The pseudo-code of Map function
Map(LongWritable key,Text value)
for each line in value do
(user,item,score)=split(line)
write(user,(item,score))
end for
The pseudo-code of Reduce function
Reduce(Text key,Iterable<TextPair> value)
Double similarity=0;
Double temp=0;
for each pair in value do
Double

$$fSim=sim(pair,currentPair)\times\frac{1}{\log(10+num)}$$

temp=temp+fSim
end for
similarity=temp+age(key,currentUser)
write(key,similarity)

FIGURE I. PSEUDO CODE OF SIMILARITY ALGORITHM ON HADOOP

V. Experiment Evaluation

A. Experimental Data Sets and Evaluation Metrics

We select the Movielens 1m dataset as the experimental data. This dataset contains 1000000 ratings from 6040 users on 3952 movies. Movies are rated on a scale of one to five, and each user has rated at least 20 movies.

For the purpose of experiments, the datasets is divided randomly in a ratio 80:20 into training and test sets.

In this paper, the MAE is used to measure the performance of the proposed algorithm. MAE is commonly used in recommender systems as the measurement of predictive accuracy. The smaller the MAE is, the higher the predictive accuracy of algorithm is [8] . The MAE is defined as follows:

$$MAE=\frac{\sum_{i=1}^{n}|P_i-R_i|}{n} \quad (4)$$

where n is the number of items, P_i is the predicted rating on item I_i, R_i is the actual rating.

B. Experimental Results and Analysis

This paper evaluates proposed algorithm respectively from the two aspects of quality and performance of the recommendation.

1) Analysis of recommended quality

We conduct experiments on the three sampled datasets and compare the MAE of the algorithm in this paper with traditional algorithm. The number of the nearest neighbour ranges from 5 to 50. Figure 2 shows the comparison of MAE for three algorithms.

We can see from the experiment results, this paper puts forward the popularity and user characteristics based on the value of the improved similarity measure method and the calculated MAE value is smaller. It indicates that the algorithm has a certain degree of improvement in the accuracy of recommendation.

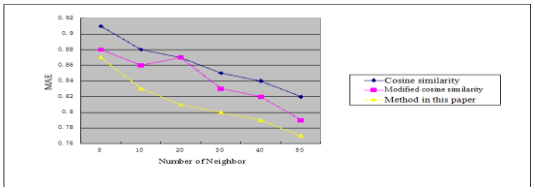

FIGURE II. COMPARISON OF MAE ON DIFFERENT ALGORITHM

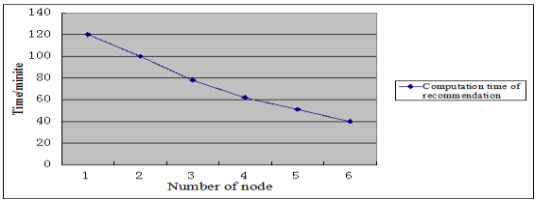

FIGURE III. THE RECOMMENDED TIME REQUIRED FOR DIFFERENT NUMBER OF NODES

2) Analysis of Recommended Performance

We use TopN approach to test the collaborative filtering recommendation algorithm based on hadoop on the MovieLens data set. That is calculation of first 10 recommended films of each user of all 6040 users in the data set. Figure 3 shows the comparison of time spent on calculating the first 10 recommended films for each user with different node (rang from 2 to 6).

Figure 3 indicates that, as the number of nodes increases, the calculation time is constantly reduced. It indicates the performance of system improved gradually.

VI. Conclusion

In this paper, by considering the relationship between the common term rating popularity and user characteristics values as two factors, we propose a new similarity measure algorithm to enhance the quality of recommendation. At the same time

the system is based on the hadoop platform. Along with the increase of the nodes, the performance could be further improved. Future work is planned to improve the Map Reduce process, such that can enable the hadoop to better deal with the algorithms.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under grants 61173192, and the Research Project of BeiLin science and technology bureau of Xi'an of China under grants GX1407.

REFERENCES

[1] VARIAN J,RESNICK P.Recommendation systems[J]. Mineapplio:Com-munications of the ACM,1997,40(3):56-58

[2] ADOMAVICIUS G,TUZHILIN A.Towards the next generation of recommender system:a survey of the state-of-the-art and possible extensions[J]. IEEE Transaction on Knowledge and Data Engineering,2005,17(6):734-749K. Elissa, "Title of paper if known," unpublished.

[3] MENG Xian-fu,CHEN Li.Collaborative filtering recommendation algorithm based on Bayesian theory[J].Journal of Computer Applications,2009,29(10): 2733-2735

[4] ZHAO Hong-xia,WANG Xin-hai,YANG Jiao-ping.Mixed collaborative recommendation algorithm based on factor analysis of user and item[J].Journal of Computer Applications, 2011,31(5):1382-1390

[5] YANG Yang,XIANG Yang,XIONG Lei .Collaborative filtering and recommendation algorithm based on matrix factorization and user nearest neighbor model[J].Journal of Computer Applications, 2012,32(2):395-398

[6] HAO Liyan,WANG Jing.Collaborative filtering recommendation algorithm based on filling and similarity confidence factor[J].Journal of Computer Applications, 2013,33(3):834-837

[7] LI Da-xue,XIE Ming-liang,ZHAO Xue-bin. Collaborative filtering recommendation algorithm based on naive Bayesian method[J].ournal of Computer Applications 2010,30(06):1523-1525

[8] Herlocker J L,Konstan J A,Terveen L G,et al.Evaluating collaborative filtering recommender system[J].ACM Transactions on Information System,2004,22(1):50-53