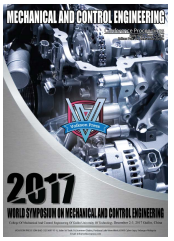
ISBN: 978-1-948012-06-5

Contents List available at VOLKSON PRESS

World Symposium on Mechanical and Control Engineering (WSMCE)



PERSONALIZED E-COMMERCE RECOMMENDATION SYSTEM BASED ON COLLABORATIVE FILTERING UNDER HADOOP

Haodong Tang, Xiaohui Cheng*

College of Information Science and Engineering, Guilin University of Technology Jiangan Road No.12, Guilin, China
Corresponding Author Email: cxiaohui@glut.edu.cn

ARTICLE DETAILS

Article History:

Received 02 october 2017
Accepted 06 october 2017
Available online 11 november 2017

Keywords

Hadoop, Collaborative filtering, personalized recommendation, implicit rating

ABSTRACT

Aiming at the disadvantages of the traditional collaborative filtering recommendation algorithm based on user rating matrix in terms of items recommendation accuracy and scalability, a personalized e-commerce recommendation system based on Hadoop is designed and implemented. The data sparseness of user-item rating matrix leads to great contingency of recommendation results, and the more sparseness scoring matrix, the worse the recommended effect. The system in this paper uses the e-commerce platform implicit user rating and explicit rating data to the user preference characteristics as the impact factor to construct and replace the rate matrix. Finally, with the modified user-item interaction information, the Top-N recommendation method is adopted to recommend the products for the target user may be interest in. Comparative experimental results show that the recommendation system can improve the recommendation accuracy of products. The experimental results on the Hadoop platform also show that using the MapReduce parallel computing framework can improve computational efficiency and algorithm scalability.

1. INTRODUCTION

After years of rapid development of the Internet e-commerce, the number of users and items on the large-scale e-commerce platform has rapidly expanded. The direct result of this phenomenon is the information overload problem. Personalized recommendation is one of the main ways to filter the huge amount of data on the Internet at present. Amazon, Alibaba and other e-commerce platforms use personalized recommendation system to recommend items to users [1]. According to VentureBeat statistics, about 30% of Amazon's sales come from its recommendation system.

Faced with such a huge number of goods on the e-commerce platform today, the user rating matrix is extremely sparse [2]. Researchers have made many improvements to solve the data sparseness, cold start, and expandability problems of traditional collaborative filtering recommendation algorithms. The null value padding is a common method of reducing the sparseness of rating matrix. Those ways use a fixed value for null padding to increase the data density.

Clustering analysis, user feature extraction and temporal context are also hot topics in collaborative filtering recommendation algorithms [3]. Starting from the e-commerce platform, this paper proposes a personalized goods recommendation system under Hadoop. The rating data is improved, and the traditional rating matrix is modified and replaced by the user's implicit rating behavior of the user on the e-commerce platform and the user's explicit rating as the influencing factor [4]. On the other hand, the Hadoop platform is used to parallelize computing and improve the efficiency and scalability of computing huge amounts of data.

2. RELATED WORK

The recommendation system consists of three modules: user modeling module, recommendation object modeling module and recommendation algorithm module, among which the recommended algorithm is the core [5]. As an excellent software framework for processing large data parallel computing, Hadoop has won the favor of big data researchers in a highly reliable, low cost and highly scalable way.

2.2 Collaborative filtering recommendation algorithm

The collaborative filtering recommendation algorithm includes item-based collaborative filtering and user-based collaborative filtering [6]. The algorithmic ideas mainly include the following three stages.

2.2.1 Construct a rating matrix

Assuming that the total number of users is m and the total number of items is n, the rating matrix R includes all user sets $U=\{u_1, u_2, \dots, u_{m-1}, u_m\}$ for all item sets $I=\{i_1, i_2, \dots, i_{n-1}, i_n\}$ score of an m × n matrix, as shown in equation (1) below.

$$R = \begin{pmatrix} r_{1,1} & r_{1,2} & \cdots & \cdots & r_{1,n-1} & r_{1,n} \\ r_{2,1} & r_{2,2} & \cdots & \cdots & r_{2,n-1} & r_{2,n} \\ \vdots & \vdots & \ddots & & \vdots & \vdots \\ \vdots & \vdots & & \ddots & \vdots & \vdots \\ r_{m-1,1} & r_{m-1,2} & \cdots & \cdots & r_{m-1,n-1} & r_{m-1,n} \\ r_{m,1} & r_{m,2} & \cdots & \vdots & r_{m,n-1} & r_{m,n} \end{pmatrix} \quad (1)$$

Wherein, each item $r_{i,j}$ in the matrix represents user i rating on item j, the non-rated item is padding by zero, and the non-zero item is the user's actual rating data on the item.

2.2.2 Calculate the similarity

The method to calculate the similarity between users or items has cosine similarity, modified cosine similarity and Pearson correlation coefficient. The cosine similarity is calculated according to the cosine of the angle between the multi-dimensional space vectors in the user-item rating matrix R, and the similarity between the user rating vector i and j is represented by Sim (i, j), as follows formula (2).

$$Sim(i,j) = \cos(i,j) = \frac{i \bullet j}{||i|| \times ||j||} = \frac{\sum_{u=1}^{m} r_{u,i} \times r_{u,j}}{\sqrt{\sum_{u=1}^{m} r_{u,i}^2} \sqrt{\sum_{u=1}^{m} r_{u,j}^2}} \quad (2)$$

Numerator is the inner product of these vectors. Denominator is the product of two vector modules. $r_{u,i}$ and $r_{u,j}$ are the ratings of user i and j in matrix respectively.

The modified cosine similarity takes into account the user rating scale. The calculation formula is shown in the following formula (3).

$$Sim(i,j) = \frac{\sum_{u \in U_{ij}} (r_{u,i} - \overline{r_u}) \times (r_{u,j} - \overline{r_u})}{\sqrt{\sum_{u \in U_i} (r_{u,i} - \overline{r_u})^2} \sqrt{\sum_{u \in U_{ij}} (r_{u,j} - \overline{r_j})^2}} \quad (3)$$

Where $U_{i,j}$ is the set of users that have a common rating for items i and j, U_i and U_j are the user sets rating for items i and j respectively, $\overline{r_u}$ is the average of all the scores of user u.

Pearson correlation coefficient is also used to calculate the correlation between two vectors, the calculation formula is as follows (4).

$$Sim(i,j) = \frac{\sum_{u \in U_{ij}} (r_{u,i} - \overline{r_u}) \times (r_{u,j} - \overline{r_j})}{\sqrt{\sum_{u \in U_{ij}} (r_{u,i} - \overline{r_i})^2} \sqrt{\sum_{u \in U_{ij}} (r_{u,i} - \overline{r_j})^2}} \quad (4)$$

2.2.3 Get the recommended list

Top-N selection strategy is a commonly used strategy for obtaining a recommendation list. Generate a list of recommendations to the target user for items that are most loved by the nearest neighbor and have not been noticed or purchased by the target user.

2.3 Hadoop big data processing

As the most widely used big data processing technology, Hadoop has two core parts: the HDFS distributed file system and the MapReduce parallel programming model.

2.3.1 HDFS

Today, large data files can reach TB or even PB level. By segmenting large files into multiple blocks and deploying multiple copies on separate, inexpensive devices, HDFS will get very good performance. Data is read and processed locally will reduce data access time and ease the computational burden on a single system.

2.3.2 MapReduce

MapReduce improves the overall computational efficiency by distributing data computing tasks to multiple nodes in a Hadoop cluster for parallel computing. MapReduce is divided into Map and Reduce; a computing task is divided into multiple Map and Reduce tasks to compute the final result in parallel.

3. COLLABORATIVE FILTERING RECOMMENDATION ALGORITHM BASED ON IMPROVED USER RATING DATA

The sparseness of user-item rating matrix makes the accuracy of the proposed results affected when cosine similarity is used to calculate item similarity. At present, some e-commerce companies maximize the user's rating data by providing bonuses or points. And some websites use user action records to recommend. Those methods have played a certain role in the improvement of data sparseness, but there are still some shortcomings.

This paper proposes a collaborative filtering recommendation algorithm based on improved user rating data. The user's implicit operation ratings and explicit ratings based on user preference characteristics together to form a new user-item rating matrix. The time factor is introduced to replace the value of the rating matrix with the new value of the user behavior, so as to more objectively and truly reflect the user's degree of love and dislike of the items. On the cold start problem, for new items, the use of site-top display to increase the probability of being accessed by the user.

3.2 New Combined User Score Matrix

We set the user action behaviors as follows: click, search, enshrine, add to the shopping cart, and purchase five operations in our recommendation system. The corresponding weights of each operation are shown in Table 1.

Table 1: User operation corresponding weight table

user behavior ($U_{operate}$)	Weight (W_o)
Click (C)	1
Search (S)	2
Enshrine (E)	3
Add to Shopping Cart (A)	4
Purchase (P)	5

The weight of the click is set to a minimum value to minimize the error introduced by the user click operation error. Other behavioral weights are in line with website operating practices. Define user actions as $U_{operate}$, the weight is W_o, as shown in the following formula (5)

.

$$w_o = \begin{cases} 1, & U_{operate} = C \\ 2, & U_{operate} = S \\ 3, & U_{operate} = E \\ 4, & U_{operate} = A \\ 5, & U_{operate} = P \end{cases} \quad (5)$$

We define the variable U_{score} as the user's score from 1 to 5 like Tmall and JDcom. The total number of rating levels expressed in $U_{quantity}$. The user preference U_p is added to the system and its value is calculated using the following formula (6).

$$U_p = \frac{U_{score}}{U_{quantity}} \quad (6)$$

Obtain the implicit user-item rating matrix through the user implicit behaviors. After user rate the products, the user's new rating for this item is calculated by following formula (7).

$$U_{u,i} = U_{u,i} \times U_p \quad (7)$$

3.3 Cold start solution

Aiming at the problems of cold start of new items and new users in collaborative filtering recommendation algorithm, we use the Mall Home display and highlighting methods to increase the probability of new products being accessed when they first enter the store Recommended for new users can be recommended hot selling items.

3.4 MapReduce process

Using the new user rating matrix as input, cosine similarity is used to calculate the similarity between the items to get the Top-N recommendation list. MapReduce process can be divided into the following four steps.

(1) Get user-item matrix by handling user behavior data format.
(2) Calculate the similarity between the various items.
(3) Calculate the user's forecast score on the item.
(4) Get Top-N recommendations list.

4. EXPERIMENT AND RESULT ANALYSIS

4.2 Experimental platform and data set

The experimental environment of this paper is using JDK 1.8.0_131 and four Hadoop 1.2.1 clusters with Centos 5.8 operating system. One of them is Master and the other three are Slaves. GroupLens team MovieLens 100k will be used as the data set, including 943 users of 1682 movies of 100000 ratings, each user rates at least 20 movies, the rating value of 1 for the particle size from 1 to 5 points is divided into 5 rating levels. The density of the rating data is 100000/(943*1682)=6.3%, the sparseness of the rating data available is 1-6.3%=93.7%.

4.3 Evaluation standard

In order to verify the quality of the recommended algorithm, the precision and the speedup is used in this recommendation system.

4.3.1 Precision

The precision is a commonly used evaluation standard to verify whether

the recommended result of the recommendation system is accurate. Since the test set does not include all users in the training set, so those users who do not exist in test sets should be excluded from the precision calculation. Because of some of the items in the recommended list also do not exist in the test dataset. So, this part of the data also will be excluded. The recommended list obtained by the algorithm is compared with the test set. Define that the score of the test set is higher than or equal to 3 is interest to the user. Otherwise, it is defined as not interested. Thus, the precision is defined as shown in equation (8) below.

$$precision = \frac{hits}{N} \qquad (8)$$

Where N is the total number of products recommended by the target user, and hits is a collection of recommended items that the target user is interested in in the recommended list.

4.3.2 Speedup

The speedup is used to verify the parallel computing performance of Hadoop clusters, as shown in Equation (9) below.

$$speedup = \frac{T_{(1)}}{T_{(N)}} \qquad (9)$$

Among them, $T_{(1)}$ stands for stand-alone running time, $T_{(n)}$ stands for n machine running time.

4.4 Result analysis

The algorithm in this paper and the traditional collaborative filtering recommendation algorithm will be test when the N were taken 5,10,15 and 20 respectively in Top-N select strategy. The algorithm of this paper will add user preference factor by the rating of 1, 2 minus 1 and the rating of 4, 5 plus 1. The experimental results are shown in Figure 1 below.

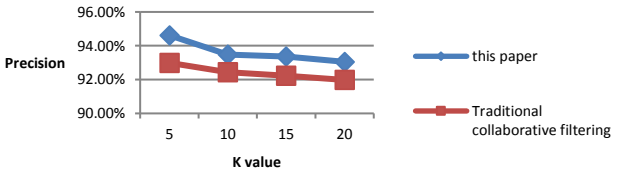


Figure 1: Precision

In order to test the parallel computing performance of the proposed algorithm in Hadoop cluster, the training data was run on ml-100k datasets on one, two and three Hadoop nodes respectively. The speedup is shown in figure 2 as follow.

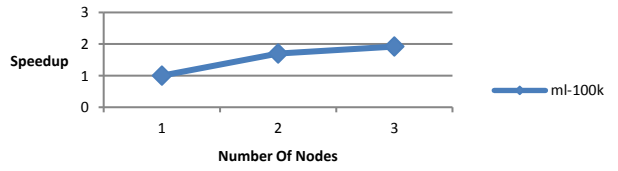


Figure 2: Speedup

As can be seen from the speedup line chart in Figure 2, the speedup increases as the number of nodes increases, which verifies that the Hadoop platform can effectively improve the operating efficiency and shorten the computation time of the recommended system.

5. CONCLUSION

In this paper, aiming at the sparseness of traditional user rating matrix, an improved approach is proposed to reconstruct the rating matrix based on the combination of user implicit behavior and rating matrix. The experimental results show that the proposed system is more accurate than traditional item-based collaborative filtering recommendation. New user-item rating matrix can better reflect the user's real interests. The algorithm runs directly on the Hadoop platform, allowing this recommendation system to have better performance in scalability and computation time compared to stand-alone systems. The experimental results show that the personalized e-commerce recommendation system based on collaborative filtering under Hadoop can predict the user's interest better and improve the accuracy of recommendation.

ACKNOWLEDGEMENTS

As the research of the thesis is sponsored by Guilin Science and Technolo gy Project Fund (No2016010408),major scientific research project of Guangxi higher edu cation (No: 201201ZD012), and Guangxi Graduate Innovation Project (No : SS201607),we would like to extend our sincere gratitude to them.

REFERENCES

[1] Gao, M., Jiang, F., Wu, Z. 2011. User rank for item-based collaborative filtering algorithm based on classification [J]. Information Processing Letters, 111 (9), 440-449.

[2] Sun, G., Shuo, W. 2015. Computing adaptive fast recommendation algorithm for user interest drift [J]. Application Research of Computers, 32 (9), 2669-2673.

[3] Wang, G., Liu, H. 2012. Review of personalized recommendation system [J]. Computer Engineering and Applications, 48 (7), 66-76.

[4] Shi, P., Zhou, Z., Wang, G. 2013. Co-Filtration Algorithm Based on Rating Matrix Pre-Filtering [J]. Computer Engineering, 39 (1), 175-182.

[5] Dong, W., Liu, D. 2017. Cooperative Filtering Recommendation Algorithm Based on Federated Clustering and User Feature Extraction [J]. Intelligence Journal, 36 (8), 852-858.

[6] An, Z., Gao, C., Xu, X.Q. 2017. Collaborative Filtering Algorithm Based on Fusion of Time and User Rating [J]. Computer Science, 44 (9), 243-249.