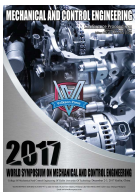
ISBN: 978-1-948012-06-5

Contents List available at VOLKSON PRESS

World Symposium on Mechanical and Control Engineering (WSMCE)



AN IMPROVED HYBRID RECOMMENDATION ALGORITHM STUDY AND DESIGN ON MAHOUT FRAMEWORK

Dai Fei, 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

Collaborative filtering algorithm, Mahout, singular value decomposition, hybrid recommendation algorithm

ABSTRACT

Collaborative filtering algorithm is one of the most widely used algorithms in the recommendation system. However, each collaborative filtering algorithm has its advantages and disadvantages. When applied to recommendation systems alone, it leads to low efficiency and low accuracy. Based on the research of existent recommendation algorithms, we improved a hybrid recommendation algorithm using Mahout Framework. This improved algorithm consists of third parts, first part: recommended results were got by using a user-based collaborative filtering algorithm to compute; second part: recommended results were got by using the improved item-based collaborative filtering algorithm to compute; third part: the two recommendation results were tested by the screening mechanism, then output the best recommended results. This paper's innovation point is that we use singular value decomposition and principal component analysis algorithm to optimize dimension reduction on item-based collaborative filtering algorithm and get better effect of the recommend system. Compared with the originally experimental results, recommendation system of the improved hybrid recommendation algorithm can have more precise outcome and high efficiency.

1. INTRODUCTION

During the Internet era, not only the electronic business platform but also information retrieval platforms and other websites are facing the issue of information overload, which leads to the issue that users cannot quickly find the information they want on the Internet, thus wasting a lot of time. The occurrence of recommendation system is of great help to solve this kind of problems. The currently popular recommendation algorithms used in recommendation system are as follows: collaborative filtering algorithm, content-based recommendation algorithm, singular value decomposition recommendation algorithm, correlation rule based recommendation algorithm, knowledge based recommendation algorithm, etc. The above algorithms have their own merits and drawbacks, and each algorithm has different applications in different scenarios.

Apache Mahout is an open source project. It not only implements most of the above recommendation algorithms in the Hadoop platform but also includes data mining, machine learning algorithms for the purpose that helps developers to quickly create a high performance intelligent application development. In China, Mahout was used in book recommendation, public search, scenic spots recommendation, online education, online news and applied in other fields widely. In view of the advantages and disadvantages of the above recommendation algorithms, many scholars have put forward the corresponding improved algorithms for the Mahout framework: Swati Pandey proposed a hybrid collaborative filtering algorithm, this method invokes user collaborative filtering and collaborative filtering algorithm based on mahout and according to the data set calculated respective results, then use a threshold mechanism to judge whether to merge the two group of results or direct gives the calculated result of item-based CF, which gets higher accuracy [1]. A researcher Proposed a new data model based on user preferences to improve the accuracy of item based recommendation algorithm [2]. A group of researchers also has combine two algorithms to ensure high accurate recommendation and high reliability [3]. In other studied, some researcher proposes an efficient multi-criteria CF algorithm, which uses dimensionality reduction techniques to improve the recommendation quality and prediction accuracy [4]. A scientist presented a hybrid recommendation method, which combines Top N Algorithm Item-based CF and MapReduce [5].

Based on the foundation, this paper gives improved method to its original hybrid recommendation algorithm [1]. That is to say, the collaborative filtering algorithm of Item-based, the core algorithm of the hybrid recommendation algorithm, is improved and optimized. The innovation point of this paper is that we use SVD and PCA to implement dimensionality reduction on the Item-based recommendation algorithm in the hybrid recommendation algorithm. Compared with original experiment result, our recommendation system can get more precise result and with high efficiency.

This paper is organized as follows: Section 2 presents the related technology of the improved algorithm. Section 3 introduces the improved hybrid recommendation algorithm in detail. We illustrate the experiment results, and evaluates the recommendation effect of the improved algorithm in section 4. At last, section 5 concludes this paper.

2. RELATED TECHNOLOGY

2.1 Singular value decomposition (SVD)

An important property of singular value decomposition is that it can provide original matrix A's best low rank approximation, which is of great use in recommendation system. Singular value decomposition (SVD) is a kind of decomposition method that can be applied to any matrix, SVD is the matrix m×n which has a follows form:

$$SVD(A) = U \sum V^T \quad (1)$$

V is a n×n orthogonal matrix,U is a m×m orthogonal matrix,$\sum$ is the m×n diagonal matrix and its diagonal elements become singular values. The columns of V consist of a set of base vectors for the orthogonal "input" or "analysis" of A. These vectors are characteristic vectors of A 、 A^T. The column of U constitutes a set of base vectors for the orthogonal output of A. These vectors are characteristic vectors of A. The elements on the $\sum$ diagonal are singular values, which can be regarded as the "expansion control" of the scalar quantity between the input and the output. These are the nonzero square roots of the eigenvalues of A 、 A^T and A^T、 A, and corresponding to the row vectors of U and V. The truncated singular value

decomposition of rank K can be defined as:

$$SVD(A_k) = U_k \sum kV_k^T \quad (2)$$

U_k and V_k are composed of M × k and N × k respectively by the matrix U Column K and the K column matrix of the V matrix, the matrix $\sum$ K is the main diagonal submatrix $\sum$ of K × K. A_k represents the nearest linear approximation of the original matrix and the downward rank K. Once the transformation is completed, users and items can be considered as points in the K dimensional space.

2.2 Principal component analysis (PCA)

The principle of principal component analysis is to try to reassemble the original variables into a new set of several independent variables that are independent of each other. At the same time, a few statistical methods that can extract as few variables as possible and reflect the information of the original variables. By means of an orthogonal transformation, it converts the original stochastic vector related to its components into a new random vector whose components are uncorrelated. It is shown in algebra that the covariance matrix of the original random vector is transformed into a diagonal matrix, and in geometry it transforms the original coordinate system into a new orthogonal coordinate system, so make it point to the Porthogonal directions that the sample points scatter most. Then the dimension reduction of multidimensional variable system is processed so that it can be converted into a low dimensional variable system with a high precision. By constructing appropriate value functions, the low dimensional system is further transformed into one dimensional system.
In this paper, the principal component analysis algorithm steps:

Step 1: Get the data from m × n matrix A;
Step 2: Calculate the covariance matrix;
Step 3: Calculate the eigenvectors and eigenvalues of the covariance matrix;
Step 4: Choosing principal components and forming a feature vector;
Step 5: Deriving the new data set and forming the clusters.

3. IMPROVED HYBRID RECOMMENDATION ALGORITHM DESIGN

The improved hybrid recommendation algorithm includes three parts: (1) user-based collaborative filtering algorithm (2) improved item-based collaborative filtering algorithm (3) after the screening mechanism and output the best recommendation result.

3.1 System architecture

Figure 1 is a method architecture diagram of the recommendation system. The three part of the hybrid recommendation algorithm in recommendation systems is described in detail in section 3.2., section 3.3., section 3.4.

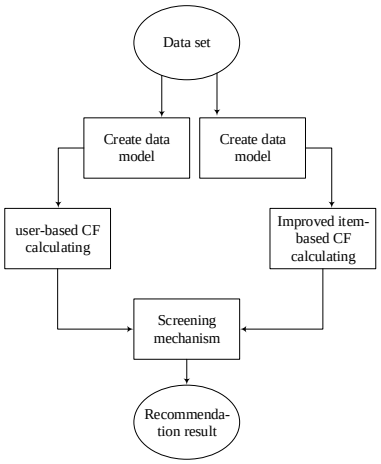


Figure 1: System framework

3.2 Calculation part of improved item-based collaborative filtering algorithm

The item-based collaborative filtering algorithm using the Singular Value Decomposition Method (SVD) not only yields accurate results but also reduces computational cost. The potential limitation is that the results may not be accurate and can not be applied directly to the 3D user-item evaluation matrix. In order to improve this shortcoming, we introduce the principal component analysis (PCA) into the hybrid recommendation algorithm. The principal component analysis (PCA) can find the linear projection of high-dimensional data to low-dimensional subspaces, for example, maximizing the variance preservation and minimizing the least squares reconstruction error. The principal component analysis (PCA) ranks input data according to the rank K of the SVD in descending order of importance and relevance.

In this way, the most important and irrelevant input components have a higher priority than the less important and highly relevant components. Input contexts that are heavily calculated are usually sparse scoring matrices, so sparse primitive inputs use PCA and SVD algorithms to increase efficiency. Later, it's easier to find potential, accurate recommendations from users. PCA is used in the calculation of the user item rating matrix missing values and for classifying and synthetizing scoring matrix. The improved item-based collaborative filtering algorithm uses the SVD and PCA algorithm to reduce the dimension of the singular value matrix, and then through the classification and synthesis, the similarity is calculated to get the similarity and score of each item.

This section of the recommendation algorithm calculation steps:

The first step: create File Data model and read the dataset into Hadoop's HDFS.

The second step: read out the data set from Hadoop HDFS and convert it into a singular value matrix, and then use the singular value decomposition and principal component analysis to decompose the singular value matrix into U、$\sum$、V. Because the scoring matrices are clustered by using the new dimension in the PCA calculation phase, the clusters of related data sets are obtained through the dimensionality reduction steps. The specific algorithm shows in section 2.1., section 2.2..
The third step: after clustering, calculate the total rating of users and items. The matrix needs to be analyzed and resynthesized to filter the singular terms of the noise data matrix. To do this, use fuzzy rules to calculate grade predictions.

The fourth step: using the Euclidean distance to calculate the similarity of the matrix, the generated distance can be converted into value of similarity.
The fifth step: after the synthesis method and similarity evaluation of the noise data from the article matrix filtering, get the list of recommended results.

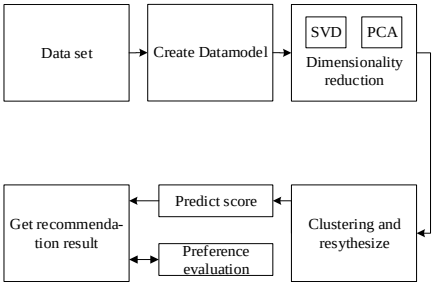


Figure 2: Improved user-based collective filtering algorithm design diagram

3.3 Calculation part of user-based collaborative filtering algorithm

This section of the recommended algorithm calculation steps:
The first step: read the data set into Hadoop's HDFS.
The second step: create File Data model calculation data model.
The third step: using Euclidean distance similarity algorithm for similarity calculation.
The fourth step: according to the neighbor model, calculating the average neighbor weights. Finally, giving a recommendation based on the preference score.
The specific process is shown in figure 3.

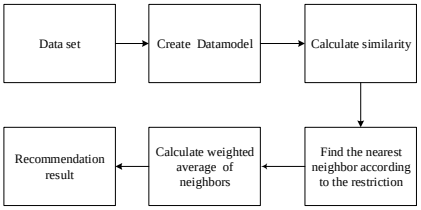


Figure 3: User-based collective filtering algorithm processing diagram

3.4 The screening mechanism that merges the recommended results

After two sets of recommendation results respectively calculated by user-based and improved item-based collaborative filtering algorithm, here comes the screening mechanism part. The process of the merger flow chart shown in figure 4.

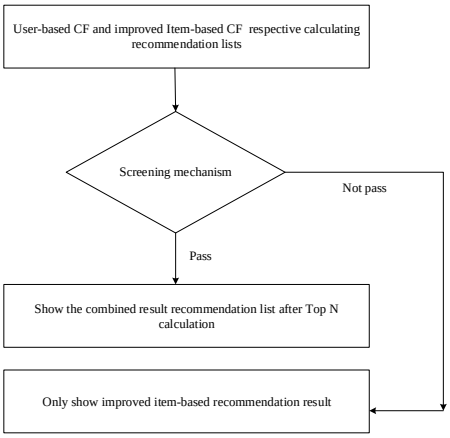


Figure 4: Computing result combination process diagram

3.4.1 Screening mechanism [1]

The screening mechanism is that compares the user-based preference value with the threshold, and if the preference value is greater than the threshold, the two recommendation results are combined and displayed after Top N algorithm calculates; if the preference value is less than the threshold value which means do not pass the screening mechanism, then only item-based recommendation results are displayed. The threshold represents the minimum number of users who prefer to recommend items, and the size of the threshold varies with the number of nearest neighbors of the item. When the threshold is raised, the accuracy of the representative items is also improved, indicating that the number of recommended suitable items also increases because more users like the item. At the same time, as the threshold increases, the number of recommended items will decrease. It is relevant and vital even if the threshold is small. Therefore, the item can directly recommend to the user after screening mechanism without additional treatment. Figure 5 shows a threshold based comparison chart.

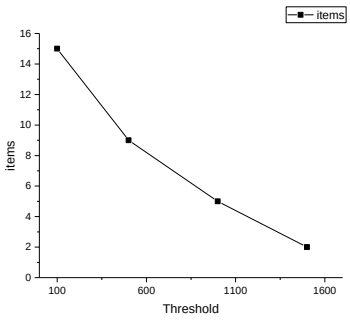


Figure 5: Threshold based comparison diagram[1]

4. EXPERIMENT AND ANALYSIS

The experimental hardware environment is AMD 3.1Ghz 4-cores processor, 8G memory machine. Based on the hardware environment, the pseudo-distributed experimental is built by virtualizing three Centos6.5 on VMware, corresponding to three nodes respectively. The data set in this paper uses Movie Lens' 100K data set, which gives 100,000 ratings (1-5 points) of 943 users to 1682 movies, each of which scores at least 20 movies. This data set contains at least three columns: user ID, rating, movie ID.

4.1 Experimental results analysis

In recommendation system, an important evaluation factor is that the recommended results can be displayed faster. This experiment is based on the Hadoop distributed system, and gets the result of two variables: speedup and efficiency. The results of these two variables change with the number of nodes. In addition, due to the inclusion of a screening mechanism, the number of recommended films is related to the change of threshold, speedup, efficiency ratio.

4.1.1 Speedup

Speedup is one of the measures of measurement performance, which defines the ratio of the task execution time in the single processing system and the execution time of the parallel processing system.

$$SpeedUp = T(1)/T(b) \quad (3)$$

In this paper, T (1) represents the execution time of a single Hadoop node, and T (b) represents the time that B nodes execute together. When this hybrid recommendation algorithm is run on a Hadoop platform, the speedup ratio varies with the number of nodes. The control experiment is the experiment of [1], the results of the comparison chart below.

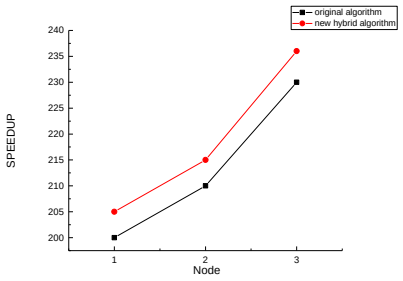


Figure 6: Speedup in different nodes

It can be seen from the figure that the speedup of the proposed algorithm is higher than that of the original algorithm, and both of them increase linearly trend.

4.1.2 Efficiency ratio

Efficiency ratio points out the usage of computing resources. It is the ratio of speedup to number of processors.

$$EfficiencyRatio = T(1)/b.T(b) \quad (4)$$

T (1) represents the execution time of a single Hadoop node, and T (b) represents the time when b nodes execute together. The efficiency ratio also represents the ratio between the processor speedup and the number of processors. When only one node is calculated, the efficiency is relatively high. As the number of nodes increases, the energy efficiency decreases. Because different nodes are used to process different tasks, the average processing efficiency increases. The efficiency of this experiment at different nodes is shown in Figure 7.

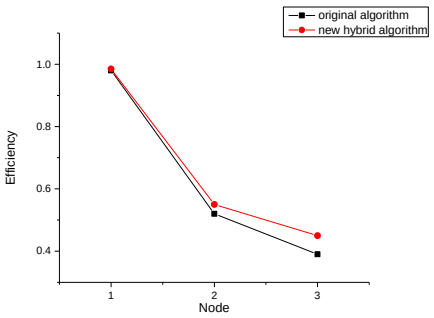


Figure 7: The efficiency of different nodes

It can be seen from the graph that the efficiency of the proposed hybrid recommendation algorithm is slightly higher than that of the original algorithm, which proves that our proposed hybrid recommendation algorithm is more optimized. The main reason is that through the SVD, PCA algorithm Dimension can reduce the computational complexity.

5. CONCLUSION

This paper presents an improved hybrid collaborative filtering algorithm based on Hadoop Mahout Framework. By improving the item-based collaborative filtering algorithm in the hybrid collaborative filtering algorithm, the accuracy of the recommended results is improved, and the computational efficiency of Hadoop is also improved. The proposed

algorithm uses SVD and PCA methods to reduce the dimensionality of item-based collaborative filtering algorithms, which not only applies to structured data, but also sparse data can be used. In the future, we plan to use more data sets to test the recommendation system or with the help of Hbase or other tools to further study in improving the algorithm.

ACKNOWLEDGMENTS

As the research of the thesis is sponsored by Guilin Science and Technology Project Fund(No：2016010408), major scientific research project of Guangxi higher education (No: 201201ZD012), and Guangxi Graduate Innovation Project (No: SS201607), we would like to extend our sincere gratitude to them.

REFERENCES

[1] Pandey, S., Kumar, S.T. 2014. Customization of Recommendation System Using Collaborative Filtering Algorithm on Cloud Using Mahout [J]. International Journal of Research in Engineering and Technology, 7 (03), 1-10.

[2] Jabakji, A., Dag, H. 2016. Improving item-based recommendation accuracy with user's preferences on Apache Mahout [J]. International Conference on Big Data, 16 (7), 1-10.

[3] Beck, P., Blaser, M., Michalke, A., Lommatzsch, A. 2017. A System for Online News Recommendations in Real-Time with Apache Mahout [J]. International Conference of the CLEF, 11 (17), 1-15.

[4] Bokde, D.K., Girase, S., Mukhopadhyay, D. 2015. An Item-Based Collaborative Filtering using Dimensionality Reduction Techniques on Mahout Framework [J]. Fourth Post Graduate Conference.

[5] Gao, X., Shi, Z. 2015. Implementation and design of the new user recommendation algorithm based on Mahout [J]. Computer science and engineering, 39 (8), 1444-1449.