Advances in Engineering Research, volume 150

4th International Conference on Machinery, Materials and Computer (MACMC 2017)

Hybrid recommendation and parallelization of movies based on spark

Mengpu Zhou[1, a], Yu Liu[2, b, *]

[1]College of Information Science and Engineering, Guilin University of Technology, Guilin 541004, China;

[2]College of Information Science and Engineering, Guilin University of Technology, Guilin 541004, China;

[a]Zmp_88@163.com, [b]Lewis_5709@163.com

Keywords: collaborative filtering, recommendation algorithm, huge data, spark.

Abstract: With the exponential growth of Internet data,the traditional stand-alone computational model has been unable to solve the real-time precise recommendation items in a complex and huge data,and the defect of traditional recommendation algorithm has become more obvious,this paper studies the collaborative filtering algorithm and matrix decomposition method,designs a parallel computing architecture based on spark,and a movie recommendation based on hybrid recommendation algorithm [1],the experimental results show that in a certain extent improves the recommendation accuracy and scalability,and has good acceleration effect.

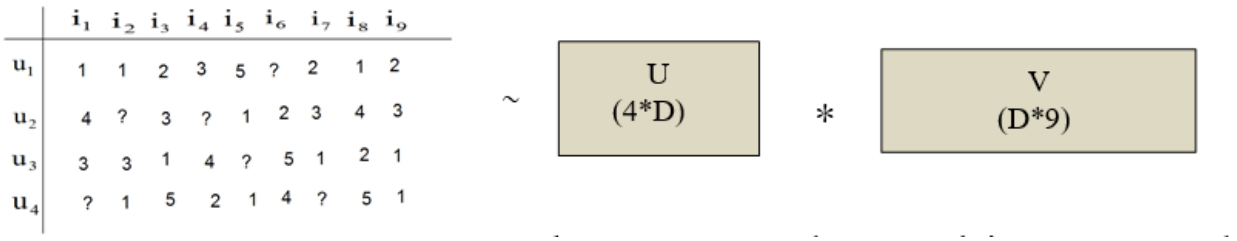
1. Introduction

When the era of rapid development of Web2.0,with the changing needs of users and the amount of data continues to expand,and the existing technical conditions,the collaborative filtering algorithm has some disadvantages [2],such as single processing efficiency is very low,a waste of computer resources,single processing mode has been unable to meet the massive data,processing speed and resource utilization are severely restricted.And the parallelism of the existing technology and processing platform is relatively low,scalability can not meet the needs of the actual business.The data sparsity and algorithm scalability lead to low accuracy,and the overall performance of the system is getting lower and lower with the increasing number of users and items.To solve the above problems,this paper studies the core idea of collaborative filtering,and the Hadoop and spark distributed computing architecture technology to the successful introduction of personalized recommendation system,and proposes a hybrid recommendation algorithm and recommendation algorithm based on hybrid spark,and passes through the strict test and repeated comparison,to a certain extent overcome recommendation the accuracy is not high,low scalability problems,and based on the parallel computing of the spark memory technology [3],improves the acceleration effect of the algorithm,greatly reduces the running time of the system.

Understand the basic algorithm and data structure,we analysis the traditional collaborative filtering recommendation algorithm and draw the following conclusions: 1) the cosine similarity matrix is normalized and improved calculation theory is the key factor, matrix factorization dimensionality reduction when the optimal RMSE index is the core [4],the favorable conditions for the realization of the film the content of high precision recommendation algorithm; 2) using Master-Slave distributed cluster parallel mode, can be easily calculated task decomposition and distributed to each computing node,improve the efficiency of computer resource utilization and recommendation systems,which are easy to implement in the cluster, the price is relatively low; 3) using the latest MovieLens data in the contrast experiment,weighted hybrid film recommendation algorithm in this research in accuracy,scalability and efficiency are excellent Improvement and improvement.

In our real life,the user item matrix is actually a great number of users and items, but for a single user, his interest and consumption ability is limited, but also cause the score recorded on the Internet

articles is also very small, very little real score data caused by the user item matrix contains a large number of null data is very sparse.



Each user represented by a D-dim vector Each item represented by a D-dim vector

Fig. 1 Matrix factorization

2. Research on hybrid algorithm

2.1 Algorithm research

According to the actual needs of personalized movie recommendation system, based on the user model (User), (Model) [5],collaborative filtering algorithm and Hadoop,spark in the environment of distributed cluster environment,study the multi-level parallel hybrid recommendation algorithm,the idea of hybrid recommendation algorithm:

A) the score matrix is normalized by scoring;

B) for the user based collaborative filtering algorithm, the improved cosine similarity is used to calculate the prediction similarity;

C) matrix reduction based on SVD;

D) RMSE index is introduced into U (user matrix) and V (feature matrix) reconstruction and decomposition to determine the optimal;

E) linear combination of the two forecast results;

F) design and implement a hybrid recommendation algorithm on spark;

2.2 Algorithm calculation

Cosine similarity without considering the user rating scale, as in the case of [1-5] score, a score of more than 3 of users is their love,and for the user B, score above 4 is your love. By subtracting the average score of the user pairs, the modified cosine similarity measure improves the above problems. The vector $\overrightarrow{R_i}$ and $\overrightarrow{R_j}$ denote the mean of the user i and the user j in the n-dimensional space respectively [6], and the similarity between the user i and the user j is :

$$sim(i,j) = \cos(i - \overrightarrow{R_i}, j - \overrightarrow{R_j}) = \frac{(i - \overrightarrow{R_i}) * (j - \overrightarrow{R_j})}{|| i - \overrightarrow{R_i} || * || j - \overrightarrow{R_j} ||} \quad (1)$$

Based on the similarity index,the nearest neighbor of the target user can be obtained, and then the recommendation result of the target user can be predicted.Let N_u^k be the set of the nearest K neighbors of the user u, and then the user u predicts the item j by : $R_{u,i}$

$$R_{u,i} = \overline{R_i} + \frac{\sum_{v \in N_u^k} sim(u,v)\,(R_{v,i} - \overline{R_v})}{\sum_{v \in N_u^k} sim(u,v)} \quad (2)$$

The deep research of UV decomposition,the film as an example,most users will only react to the characteristics of small scale,they love some schools,may love some famous movie stars,or some love director works who has many followers,if the M utility matrix with n rows and m columns, so we can find a n matrix U D column and a d m column of the V matrix, which makes the UV and M in non empty elements M are very similar. If so,we can confirm the D feature that allows us to characterize

and get close the user, you can use the UV to estimate the elements of the corresponding blank elements utility matrix in M.

As before, suppose that M is a utility matrix of n*m with some blank elements, while U and V are matrices of n*d and d*m dimensions respectively. We use m_{ij}, u_{ij} and v_{ij} to represent the elements of column i, row j, M, U and V respectively. In addition, we assume P=UV, and use p_{ij} to represent the elements of column i, column j in product matrix P. Suppose we change the u_{rs} to find the minimum element values that make RMSE between M and UV minimum. Notice that u_{rs} only affects elements of line r of P=UV. So, we only need to focus on all the j values (as long as m_{rj} is not empty):

$$P_{rj} = \sum_{k=1}^{d} u_{rk}V_{kj} = \sum_{k\neq s} u_{rk}V_{kj} + xV_{sj} \quad (3)$$

In the above expression, we have replaced the element u_{rs} that we want to change into variable x, and used a convention expression, That is $\sum_{k\neq s}$ expresses the summation result of k=1,2,...,d except k=s. If m_{rj} is a non empty element in the matrix M, then the element contributes to the sum of squared errors:

$$(m_{rj} - p_{rj})^2 = (m_{rj} - \sum_{k\neq s} u_{rk}V_{kj} - xV_{sj})^2 \quad (4)$$

Here we use another convention expression, that is $\sum_{k\neq s}$ means the sum of all the j when m_{rj} is nonempty. So we can write the sum of squares of all errors affected by x= u_{rs} as an expression:

$$\sum_{j}(m_{rj}-)\sum_{k\neq s}(u_{rk}V_{kj} - xV_{sj})^2 \quad (5)$$

Here, we use the root mean squared error (RMSE) as the evaluation index:

$$RMSE= \sqrt{\frac{1}{|D^p|}\sum_{(u,\alpha)\in E^p}(r_{u\alpha} - r'_{u\alpha})^2} \quad (6)$$

In order to obtain the minimum x of RMSE, the above formula is derivative to X and the other is 0:

$$\sum_{j} -2V_{sj}(m_{rj} - \sum_{k\neq s} u_{rk}V_{kj} - xV_{sj}) = 0 \quad (7)$$

As with the previous example, we can ignore the constant factor -2 and solve the above equation about x:

$$x = \frac{\sum_{j} V_{sj}(m_{rj} - \sum_{k\neq s} u_{rk}V_{kj})}{\sum_{j} V_{sj}^2} \quad (8)$$

For an element of V, there is a similar formula for its optimal value.If we want to change the A, then make the minimum y value of RMSE be:

$$y = \frac{\sum_{i} u_{ir}(m_{is} - \sum_{k\neq r} u_{ik}V_{ks})}{\sum_{i} u_{ir}^2} \quad (9)$$

Here, it represents the sum on all i when B is nonempty, and C represents the summation result of k=1,2,...,d except k=r. Iterations are constantly carried out to obtain the best element values [7].

Calculation of mean square error RMSE of definition function [8]:

```
Def computeRmse(model：MatrixFactorizationModel,data:RDD[Rating]):
Double={
```

```
    val predictions:RDD[Rating]=model.predict(data.map(x=>(x.user,x.product)))
valpredictionsAndRatings=predictions.map{x=>((x.user,x.product),x.rating)}.join(data.map(x=>((x
.user,x.product),x.rating))).values
math.sqrt(predictionsAndRatings.map(x=>(x._1-x._2)*(x._1-x._2)).mean())
}
```

3. Experiment and evaluation

Hadoop is a software framework for distributed processing of large amounts of data. It can process massive data in a reliable, efficient and scalable way. The core structure of Hadoop is HDFS, MapReduce and Yarn. But the real-time processing effect of Hadoop is poor, and spark is compatible with HDFS and Hive distributed storage layer, to a certain extent make up for the defects of MapReduce [9], and it is not only based on memory computing, but also has a strong advantage in the data format, strategy, execution task scheduling.

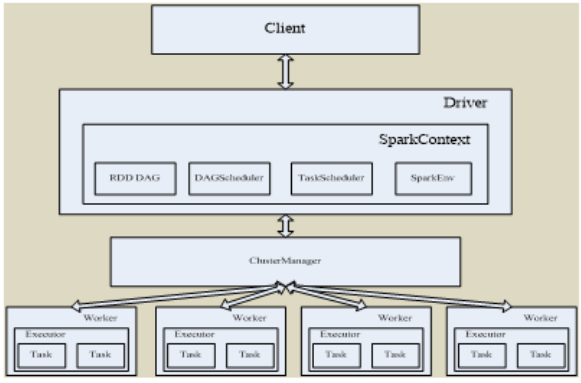


Fig. 2 Source processing architecture

Spark is a superior performance of the distributed computing framework.RDD is the distributed memory data abstraction, also known as flexible distributed data sets, and a lot of machine data were distinguished, RDD comprised of multiple partition, according to the number of which is cut into pieces, then each partition can correspond to a task, eventually reaching the effect of parallel computing [10].

Laboratory environment:

1) Four nodes PC cluster test connection in 1000Mbps network, each node configuration 64 CentOS 6.4 operating system, Intel (R) Core (TM) i5-3470 CPU (8GB, GeForce GTX Kepler 680GPU memory architecture, stream processor 1536). Cluster installation of Hadoop2.7.0, Spark2.11.8, Scala2.12.2, JDK1.7.0.

2) Experiment for the system efficiency: the two variables in the experiment, the number of nodes and the amount of data,we do the experiment with the number of nodes with different amount of user data, validate the algorithm and system speedup effect, which uses 2 nodes, 3 nodes and 4 nodes respectively, the amount of data from 2000,4000,6000,8000,10000 followed by experiment. The experiment results show that with the increase of user data, the number of nodes in the cluster increases, the running time of the whole system is reduced, and the acceleration effect is obvious. And in the condition that the RMSE evaluation index is basically stable, the recommendation system of this paper has better operation efficiency, and has certain improvement in scalability.

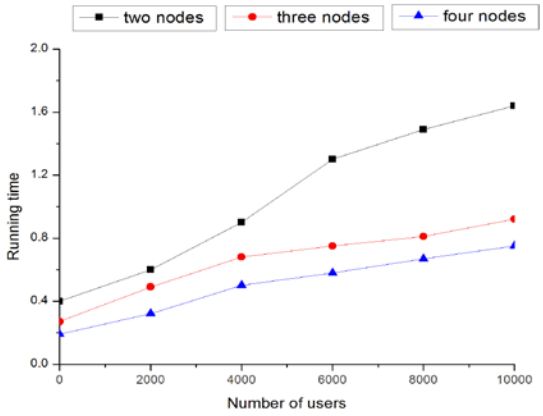


Fig. 3 Experimental results table

4. Summary

Through in-depth study of the traditional collaborative filtering algorithm,and proposes a hybrid recommendation algorithm,the purpose is to improve the effectiveness of the recommendation.We used the Hadoop distributed data processing platform recommendation algorithm,and the introduction of spark core technology,aiming to improve the data processing speed of the recommendation system and its scalability through the experiment.The recommendation system based on the distributed system has a certain acceleration effect [11]. With the increasing amount of data,the scalability of the algorithm is also improved.In many practical projects,effective evaluation data is generally below 1%,and even in some cases will be less.The process of dimensionality reduction by matrix decomposition,to a certain extent,improve the user similarity accuracy.Thereby improving the accuracy of the personalized recommendation system,has a good user the effect of experience.In the following work,a larger amount of data will be used to optimize the recommendation algorithm theory,and maximize the recommendation effect of the distributed recommendation system.

Acknowledgments

The work reported in this paper has been supported by The National Natural Science Foundation of China under research project 41264005 . We would also like to thank the anonymous reviewers for their comments and valuable suggestions.

References

[1] Shouxian Wei,Xiaolin Zheng,Deren Chen,Chaochao Chen. A hybrid approach for movie recommendation via tags and ratings[J]. Electronic Commerce Research and Applications,2016.

[2] Feng Ge. A Collaborative Filtering Recommendation Approach Based on User Rating Similarity and User Attribute Similarity[J]. Advanced Materials Research,2014,2863(846).

[3] Ankur Narang,Abhinav Srivastava,Naga Praveen Kumar Katta. Distributed Scalable Collaborative Filtering Algorithm[M].Springer Berlin Heidelberg:2011.

[4] Hui Xia. Research on Recommendation Algorithm of Matrix Factorization Method Based on MapReduce[J]. Applied Mechanics and Materials,2014,3458(631).

[5] Zhao Deng,Jin Wang. Collaborative Filtering Algorithm Based on User Clustering[J]. Applied Mechanics and Materials,2013,2700(411).

[6] Yajun Leng,Qing Lu,Changyong Liang. A collaborative filtering similarity measure based on potential field[J]. Kybernetes,2016,45(3).

[7] Machine Learning; Findings on Machine Learning Detailed by Investigators at Zhejiang University (User Preference Learning for Online Social Recommendation)[J]. Journal of Robotics & Machine Learning,2016.

[8] T. Chai,R. R. Draxler. Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature[J]. Geoscientific Model Development,2014,7(3).

[9] M.Bakratsas,P.Basaras,D. Katsaros,L.Tassiulas. Hadoop MapReduce performance on SSDs for analyzing social networks[J]. Big Data Research,2017.

[10] Sasmita Panigrahi,Rakesh Ku. Lenka,Ananya Stitipragyan. A Hybrid Distributed Collaborative Filtering Recommender Engine Using Apache Spark[J]. Procedia Computer Science,2016,83.

[11] Chengxiang Si,Xiaoxuan Meng,Lu Xu. High Performance Computing Systems and Applications[M].Springer Berlin Heidelberg:2010.