A recommendation method of teaching resources based on similarity and ALS

Chuanwei Geng*, Jian Zhang, Lianzheng Guan
Department of Information, Beijing University of Technology, Beijing, China

*Corresponding author: gchw@emails.bjut.edu.cn

Abstract. This paper provides a collaborative filtering recommendation algorithm based on the combination of item similarity and alternating least square collaborative filtering based on the Spark platform. This is a recommended method to improve the efficiency of prediction calculations and reduce system response time. In order to solve the problem of model inaccuracy caused by the sparse data of the existing collaborative filtering recommendation scheme, which leads to the inaccuracy of recommending suitable online education and teaching resources to different users, the present invention uses the least squares collaborative filtering recommendation algorithm on the Spark big data analysis platform Optimize and use, and then use parallel methods to increase the amount of work completed per unit time and the accuracy of recommendations, and solve the problem of inaccurate recommendation of teaching resources.

Keywords: Education, Item Similarity, Alternating Least Squares, recommendation

1. Introduction

Under the epidemic situation, various education and teaching platforms have appeared one after another, and online education and teaching resources are also growing at a rapid rate. However, the sharp increase in the amount of education and teaching resources has brought many troubles and challenges to teaching organizers and learners. I have to spend a lot of time and energy to screen out educational and teaching resources that meet my needs. Therefore, the personalized recommendation system, which is widely used in the business field, has also begun to be gradually applied to the education and teaching field. It can use the user's historical behavior data to perform personalized calculations, discover the interests of different users, and guide users to gradually find their needs. Information and education and teaching resources, which greatly improves the work and learning efficiency of users.

Recommendation algorithm is a method that can filter information. It can effectively recommend personalized information according to users' information needs, personal interests, etc., and has been successfully applied in online videos, social platforms, online music, e-commerce, etc. field. With the continuous improvement of education and teaching resources in the construction of smart education and teaching, the use of rich education and teaching resources, such as e-books, teaching videos, documents, etc., to make personalized recommendations can help improve students' learning efficiency.

The most popular recommendation algorithm at present is Collaborative filtering (CF) [1]. Collaborative filtering, literally understood, includes two operations: collaboration and filtering. The so-called synergy is the use of group behavior to make recommendations. Biologically, there is a saying of
co-evolution. Through synergy, the group gradually evolves to a better state. For recommendation systems, through the continuous synergy of users, the final recommendations to users will become more accurate. And filtering is to find (filter) the user's favorite plan from the feasible decision (recommendation) plan. Specifically, the idea of collaborative filtering is to find a certain similarity (the similarity between users or the similarity between the subject matter) through group behavior, and make decisions and recommendations for users through the similarity. Generally speaking, collaborative filtering recommendations are divided into three types. The first is user-based collaborative filtering [2], the second is item-based collaborative filtering [3], and the third is model-based collaborative filtering. User-based collaborative filtering mainly considers the similarity between users and users. As long as you find out the educational and teaching resources that similar users like, and predict the target user’s rating of the corresponding educational and teaching resources, you can find several education and teaching resources with the highest ratings are recommended to users. And item-based collaborative filtering is similar to user-based collaborative filtering, but at this time we turn to finding the similarity between educational teaching resources and educational teaching resources, only to find the target user’s preference for certain educational teaching resources Then we can predict similar educational and teaching resources with high similarity, and recommend several similar educational and teaching resources with the highest scores to users. For example, if you buy a machine learning-related book on the Internet, the website will immediately recommend a bunch of machine learning and big data-related books to you. This obviously uses the idea of project-based collaborative filtering.

However, the accuracy of many recommendation algorithms is low, and the direct use of collaborative filtering methods will cause the recommended education and teaching resources to be out of date due to lack of understanding and understanding of the course content, and the recommendation results are less relevant to the prepared course tasks. Therefore, it is necessary to construct a personalized recommendation method that recommends relevant education and teaching resources as accurately as possible.

2. Methodology

2.1. Introduction to the Spark platform
Spark is an open source distributed cloud computing platform with file system, database, data processing system, machine learning library, and allows developers to store data in memory [4]. It has an advanced directed acyclic graph (DAG) to run the engine to support data flow and memory calculation, so the program running speed is greatly improved. Spark also provides efficient data manipulation, making the development of parallel applications very simple.

Spark has two abstractions, the first is RDD (Resilient Distributed Dataset), and the other is shared variables. RDD is one of Spark's core concepts. It represents records, which is a certain type of object. These objects will be distributed to different nodes in the same cluster. When a node or task fails, the RDD in Spark is fault-tolerant. The fault-tolerance is caused by different networks, hardware failures, etc., rather than code errors. Shared variables are variables that all nodes can share. The reason for the existence of this kind of variable is that when parallel, the function of all variables in each node makes the copy of the variable by the node itself does not affect the change of the variable of another node [5].

Spark's parallel and fast machine learning engine-MLlib (Machine Learning lib, machine learning library), including commonly used machine learning algorithms, and related data generators and tests. The performance of many iterative machine learning algorithms on Spark is very good. Currently, the main machine learning algorithms of Spark's MLlib include collaborative filtering, regression, clustering and classification. Personalized recommendation engines often use collaborative filtering recommendation algorithms because they can infer items that users might like. Today, MLlib only supports model-based collaborative filtering recommendation ALS. Its idea is to reduce the size of the matrix through matrix decomposition, taking into account the hidden factors between the product and the user. These factors help to more accurately infer user preferences [6].
2.2. Item-based Collaborative Filtering

The recommendation idea of Item-based collaborative filtering is to recommend items similar to their favorite items to target users. The similarity of items is based on the simultaneous appearance of items in user behavior records.

The ItemCF algorithm is mainly divided into two steps:

Step1: Calculate the similarity between items.

Step2: Generate a recommendation list of items that are highly similar to the items that the target user likes for the target user.

![Diagram of ItemCF algorithm](image)

**Figure 1.** The basic principle of ItemCF algorithm

The solid line indicates the existing user behavior records. In Fig. 1., User1 and User3 both prefer ItemA and ItemC, so ItemA and ItemC have a high similarity, and ItemC can be recommended to User2 who is interested in ItemA.
2.3. **Alternating Least Squares**

Alternating Least Squares [7] observes the user’s rating of the item through matrix decomposition. Based on the rating matrix is a low-rank matrix, the main idea is to find two low-dimensional matrices $U_{(m \times k)}$ and matrix $V_{(n \times k)}$ to approximate $R_{(m \times n)}$.

$$R_{(m \times n)} \approx U_{(m \times n)}V_{(n \times k)}$$  \hspace{1cm} (1)

Fig. 2. is the R matrix decomposition model diagram, which explains how to decompose the R matrix.

![Figure 2. R matrix decomposition model diagram](image)

The specific solution method of Alternating Least Squares is an alternate iterative calculation process. First, fix one of the matrices such as $V$, take the partial derivative of the loss function $L(U, V)$ with respect to $U_i$, and set the derivative equal to 0 to obtain:

$$U_i = R_iV_{u_i}(V_{u_i}^TV_{u_i} + \lambda n_{u_i}I)^{-1}$$  \hspace{1cm} (2)

Fixing $U$ in the same way, we can get:

$$V_j = R_jU_{v_j}(U_{v_j}^TU_{v_j} + \lambda n_{v_j}I)^{-1}$$  \hspace{1cm} (3)

Repeat the alternate calculation of $U_i$ and $V_j$ to update the matrices $U$ and $V$ until the root mean square error changes little or reaches the maximum number of iterations [8], and $\hat{R} = XY$ is obtained.

2.4. **Item Similarity and Alternating Least Squares**

In order to combine the advantages of the project-based similarity and the alternating least squares algorithm and make up for the shortcomings of the two, in this article, the project similarity is combined with the ALS algorithm. The flowchart is shown in Fig. 3.
Step 1: First, obtain the user’s rating matrix through the raw data of the user’s rating, and calculate the similarity between the user and the user, the educational teaching resource and the educational teaching resource. The present invention uses the vector cosine method, also called VSS (spatial similarity method) [9], where N(m) represents the set of educational and teaching resources owned by user m, that is, the promotion of educational and teaching resources by user m, N(n) Indicates the educational and teaching resources owned by user n, and the similarity formula between users m and n is:

$$PC(m, n) = \frac{|N(m) \cap N(n)|}{|N(m)||N(n)|}$$  

(4)

It can be known from the same principle that the similarity formula between educational teaching resource i and educational teaching resource j is:
\[
PC(i,j) = \frac{|N(i) \cap N(j)|}{|N(i)| |N(j)|}
\]  \hspace{1cm} (5)

Step 2: Design a loss function and add the similarity data obtained by the above calculation to the function of equation (8).

\[
f(U, V) = \sum_{(i,j) \in I} (r_{ij} - u^T_i v_j)^2 \\
+ \sum_{i \in I} \left( u_{ki} - \frac{\sum_{p \in E(K(u_i))} PC(m,p) u_{kp}}{\sum_{p \in E(K(u_i))} PC(m,p)} \right)^2 \\
+ \sum_{j \in I} \left( v_{kj} - \frac{\sum_{q \in E(K(v_j))} PC(i,q) v_{kq}}{\sum_{q \in E(K(v_j))} PC(i,q)} \right)^2 \\
+ \lambda \left( \sum_i n_{ui} \|u_i\|^2 + \sum_j n_{vj} \|v_j\|^2 \right)
\]  \hspace{1cm} (6)

Step 3. Perform iteration to solve the minimization function in the function of step 2, and finally obtain U and V matrices.

\[
\frac{1}{2} - \frac{\partial f}{\partial u_{kl}} = 0, \hspace{0.5cm} \forall i, k
\]

\[
\Rightarrow \sum_{i \in I} (u_i^T v_j - r_{ij}) v_{kj} + (u_{ki} - \frac{\sum_{p \in E(K(u_i))} PC(m,p) u_{kp}}{\sum_{p \in E(K(u_i))} PC(m,p)}) + \lambda n_{ui} u_{ki} = 0, \hspace{0.5cm} \forall j, k
\]

\[
\Rightarrow \sum_{i \in I} (v_{kj} u_i^T) + (\lambda n_{ui} + 1) u_{ki} = \sum_{i \in I} v_{kj} r_{ij} + \frac{\sum_{p \in E(K(u_i))} PC(m,p) u_{kp}}{\sum_{p \in E(K(u_i))} PC(m,p)}, \hspace{0.5cm} \forall j, k
\]

\[
\Rightarrow u_i = (V_j V_j^T + (\lambda n_{ui} + 1)E)^{-1} \left[ U_i R^T(i,i) + \frac{\sum_{p \in E(K(u_i))} PC(m,p) u_{ip}}{\sum_{p \in E(K(u_i))} PC(m,p)} \right], \hspace{0.5cm} \forall j
\]  \hspace{1cm} (7)

Solve U and V exactly, solve V according to a fixed U, solve U with a fixed V, and iteratively solve using the alternating least square method. The specific steps are as follows:

The same about \(V_j\), among them, there are:

\[
v_j = (U_i U_i^T + (\lambda n_{vj} + 1)E)^{-1} \left[ V_j R^T(j,j) + \frac{\sum_{p \in E(K(v_j))} PC(j,p) v_{pq}}{\sum_{p \in E(K(v_j))} PC(j,p)} \right], \hspace{0.5cm} \forall j
\]  \hspace{1cm} (8)

3. Data and Result Analysis

Since not data set related to teaching resources was found, this article uses the MovieLens data set provided by the GroupLens research team. The data set version selection includes the latest version of about 700 users’ ratings of 9,000 movies, and about 100,000 scoring data. The experiment environment adopts the ubuntu 16.04 virtual Hadoop cluster set up on Macbot, which contains 3 nodes in total. The storage method adopts HDFS, the Spark cluster relies on Yarn, and the programming language uses Python.

The scoring data is split into training data and test data. The training data is imported into the algorithm as a data source to calculate the prediction score, and the test data is used to verify the prediction score. In order to evaluate the accuracy of the proposed algorithm, we use the following evaluation indicators: mean square error (MSE), root mean square error (RMSE), global average
accuracy (MAP). MSE means that the vibration range of the data evaluation changes. If the range is small, the correctness of the proposed model is very good. RMSE serves as the standard deviation. MAP is a ranking based on the evaluation index. A higher value indicates a higher accuracy rate of the recommended resources.

![Figure 4. MSE result evaluation comparison chart](image1)

The evaluation result of MSE is shown in Fig. 4, and it can be known from the result of Fig. 4 that the recommendation algorithm used in the present invention is more accurate than the two compared algorithms. The evaluation result of MAP is shown in Fig5. It can be seen from Fig5 that the algorithm used in this article greatly exceeds the algorithm compared with it. MSE shows the accuracy of the algorithm proposed in this article, and MAP can show the ranking of recommended results. According
to the results, it can be seen that these two algorithms can complement each other to meet the needs of the system.

4. Conclusions
This article first introduces the Spark platform, and points out its advantages and widespread use; then analyzes the current popular collaborative filtering algorithm based on item similarity and alternating least squares algorithm, and proposes the use of an improved cosine similarity algorithm to calculate item similarity. After solving the matrix similarity degree, it is proposed to substitute it into the loss function, and then perform matrix decomposition; finally, in the Spark cluster environment, use the MovieLens data set to execute the algorithm. The experiment verifies the superiority of the algorithm through comparative analysis.

Acknowledgments
We sincerely thank the anonymous reviewers for their valuable comments and suggestions.

References
[1] Y. Xiwang, G. Yang, L. Yong, S. Harald. A survey of collaborative filtering based social recommender systems [J]. Computer Communications, 2019, 41.
[2] Y.L.Zhang, M.M.Ma, S.P. Wang. Research of user-based collaborative filtering recommendation algorithm based on hadoop [A]. Science and Engineering Research Center. Proceedings of 2015 International Conference on Computer Information Systems and Industrial Applications (CISIA2015) [C]. Science and Engineering Research Center: Science and Engineering Research Center, 2015:4.
[3] W.H.Hu, F.Yang, Z.W.Feng. Item-based collaborative filtering recommendation algorithm based on MapReduce [A]. Advanced Science and Industry Research Center. Proceedings of the International Conference on Multimedia, Communication and Computing Application(MCCA 2014)[C]. Advanced Science and Industry Research Center: Science and Engineering Research Center, 2014: 6.
[4] S. Guoxing, Z. Jiaming, Z. Qingwei, X. Shao. Design and Research of Machine Learning Platform Based on Spark Architecture [J]. Information Technology and Information Technology, 2020 (05): 55-58.
[5] Q. Rongcai. Parallel design and application of CURE algorithm based on Spark platform [D]. South China University of Technology, 2019.
[6] J. Tingting. ALS collaborative filtering recommendation algorithm research on Spark platform [D]. Dalian Maritime University, 2018.
[7] B. Mitroi, F. Frasincar. An elastic net regularized matrix factorization technique for recommender systems [P]. Applied Computing, 2020.
[8] L. Jung Bin, Szu Yin Lin, Y. Hsiang Hsu, Y. Chu Huang. An empirical study of alternating least squares collaborative filtering recommendation for Movielens on Apache Hadoop and Spark [J]. International Journal of Grid and Utility Computing, 2020, 11 (5).
[9] X. Chuping, S. Hua. A Personalized Collaborative Filtering Recommendation Algorithm Based on Linear Regression [J]. Mathematical Modelling of Engineering Problems, 2019, 6 (3)