Research on Personalized Recommendation Algorithm Based on Collaborative Filtering and Partition Clustering

Zhe WANG
College of Technology and Information, Agricultural University of HeBei, BaoDing 071000, HeBei Province, China

Keywords: Recommendation algorithm, Collaborative filtering, Partitioning clustering, User interest matrix, Neighborhood.

Abstract. Collaborative filtering is one of the most widely used and successful recommendation technologies in e-commerce recommendation system. However, the traditional collaborative filtering recommendation algorithm is confronted with the problem of data sparseness in the face of geometric multiplication of agricultural products, leading to reduced efficiency and accuracy. To solve this problem, this paper proposes a collaborative filtering recommendation algorithm based on partition clustering. In the traditional collaborative filtering algorithm, the idea of user interest matrix clustering. Firstly use the known user clustering; Secondly, the degree of the centralized user and each cluster center is obtained by the user interest matrix; then get the similarity between the target user and the known user; Eventually get recommended. The experimental results show that the improved algorithm makes the spatial search ratio of the target user's nearest neighbor search greatly reduced, and improve the efficiency and accuracy of the recommendation to a certain extent.

Introduction

With the growing development of big data, e-commerce has become more and more life choice for users. The initial simple transaction has been unable to meet the growing needs of users, targeted personalized service will be generated, along with the recommendation system has gradually improved[1-3]. At present, the most widely used algorithm has three types: First, establish a sequence by collecting and analyzing the basic information of users and daily search habits to complete the recommendation, called the sequence-based recommendation algorithm [4]; The second is to collect and analyze the online resources and the association and similarity of users to complete the recommendation, called the resource and user similarity recommended algorithm[5,6]; Third, utilized the similarity between users, filter out the items that other users like and recommended to the target user, known as collaborative filtering recommendation algorithm[7-9]. Collaborative filtering recommendation algorithm has the characteristics of high accuracy and small workload, which has become the main recommendation algorithm in e-commerce[10,11].

Related Work

Traditional Collaborative Filtering Recommendation Algorithm

The traditional collaborative filtering recommendation algorithm considers that there is a similar interest among similar users, the set of objects recommended to the target user is generated from the set of interest objects of similar users, that is, assuming that a user set \( u_i, u_{j, j \neq i} \) in \( U \) is a similar user, If the user \( u_i \) score a product for the \( R_{ui} \), the user \( u_{j, j \neq i} \) on the product rating may also be \( R_{ui} \). So it is recommended to the target user's object set does not need to consider the form of expression and specific content, only to analyze the similar user behavior data to get the recommended object, and the recommendation accuracy is relatively high. Therefore, the recommended step is to determine the similar user in the user set, and then calculate the object set to be recommended to the target user according to the behavior analysis of the similar user.

In summary, we can see that the collaborative filtering algorithm is based on the score matrix of the existing users to get the recommended object set for the target user. The recommended quality is...
closely related to the detail degree of the score matrix, that is, the more accurate the user's evaluation of the project object.

**Division Clustering Algorithm**

Cluster analysis is an effective means of data mining, in the context of unsupervised learning from a variety of cluster data set to find the internal relationship. The data are re-clustered according to the similarity of the similarity, so that the similarity of the data objects in the cluster is as large as possible, and the dissimilarity of the data between different clusters is as large as possible.

The basic steps of dividing the cluster: Starting from the partitioning of the initialization data, the original data set is divided into K clusters by optimizing an evaluation function. Secondly, the results of the partitioning are optimized. According to the relocation technology and the similarity function, the objects included in each cluster are transformed from each other until the best result is obtained. A rational division principle is that the data objects within the same cluster are maximized close to or similar to each other, and that different classes of data objects should be maximized away from each other. The classical division method commonly used is the K-medoids algorithm and the K-means (K-means) algorithm.

**Algorithm Improvement**

In this paper, based on the traditional cooperative filtering algorithm, the idea of user-interest matrix clustering is proposed to improve the recommendation accuracy.

Based on the idea of word frequency-inverse document frequency (TF-IDF) in content co-filtering, we define the degree of preference of user i for a product attribute p, as in Equation 1, and calculate the degree of preference based on user score matrix and project attribute matrix to obtain user-interest matrix.

$$\text{Interest}(m, p) = \text{TF}(m, p) \times \text{IDF}(p)$$  \hspace{1cm} (1)

$$\text{TP}(m, p) = \frac{H_{mp}}{\sum_{i=1}^{H} H_{mi}}$$  \hspace{1cm} (2)

$$\text{IDF}(p) = \log \frac{Y}{y_p}$$  \hspace{1cm} (3)

In the Equation 2, H represents the total number of attributes owned by the product, $H_{mp}$ represents the number of attributes p that the user m is interested in, and $\sum_{i=1}^{H} H_{mi}$ represents the total number of attributes that the user m has all the products of interest. $Y$ is the number of products with attribute p, and Equation 3 indicates that the attribute p in a product that a user likes is the proportion of all products with this attribute. The higher the value is, the more the user likes this attribute.

On the basis of obtaining the user - interest matrix, the algorithm implementation recommendation also needs to cluster and find the neighbor set.
In the above algorithm, the target user's neighbor recommendation. First, the user-interest matrix is used to calculate the degree of the user's centralized user and each cluster center, and then the similarity between the target user and the known user is calculated according to the user-scoring matrix, and finally the neighboring set Top K of the target user is obtained.

Finally, in the use of formula 2 to obtain the target user to not view or not rated the product score, while the score in descending order, in order to get the final recommendation of the target user Top N.

**Experiment Analysis**

**Introduce the Data Set**

The experimental data used in the current recommendation system commonly used by the University of Minnesota published MovieLens movie score data set. The data set is divided into three versions based on the film score data: one hundred thousand, one million, ten million, this experiment selected scoring data for the order of 100,000 orders of magnitude.

Experiments used in the MovieLens data set of about a thousand users of about 1682 movies about one hundred thousand anonymous score data. Which contains the film name, number, belongs to the genre and other film attribute collection; users according to personal preferences for the film 1-5 score of the film score collection and user ID and other personal information collection. In order to make the score set more complete and accurate, Minnesota University removed the score score of less than 20 films and personal information imperfect users, and to ensure that the collection of the film at least two evaluation.
Evaluation Index

This article is intended to improve the recommended real-time and accuracy. Real-time explanation for the consideration of new products to join or based on changes in user behavior to re-generate recommendations. The accuracy of the recommendation is the difference between the existing score of the product and the predicted score of the target user for the unrated product by the algorithm. The smaller the value is, the higher the accuracy is.

Reference [10] defines the spatial search rate to represent the recommended real-time, as in Equation 4; using the mean absolute deviation MAE to analyze the recommended accuracy, as in Equation 5.

\[
\text{Ratio} = \frac{\text{NU}_1}{\text{NU}_2}
\]

\[
\text{MAE} = \frac{\sum_{u,i \in T} |R_{ui} - r_{ui}|}{|T|}
\]

In Equation 4, \( \text{NU}_1 = C_1 \cup C_2 \cup \ldots \cup C_k \) is the neighborhood set obtained by the algorithm improvement, and \( \text{NU}_2 \) is the neighbor set obtained by the traditional cooperative filtering algorithm. The smaller the value is, the better the real time is suggested. In Equation 5, \( R_{ui} \) indicates that user \( u \) scores a product \( i \), \( r_{ui} \) is the algorithm predictive score, and \( T \) is the experimental test set.

Experimental Results and Analysis

Select the number of clusters were 10, 20 and 30, respectively, the film data cluster clustering, and come to the following experimental results:

![Figure 1. Algorithm to improve the spatial search rate.](image1)

![Figure 2. Comparison of recommended performance of traditional collaborative filtering and improved algorithms.](image2)

It can be seen from Figure 1, when the specified number of clusters is much smaller than the number of products within the data set, the recommended rate and the number of clusters related to the size: the larger the number of clusters, the less the number of clusters to be searched for when looking for the neighbors of the target user, thus reducing the search time. The experimental results show that the spatial search ratio of the target users is greatly reduced after the clustering idea is integrated on the basis of the collaborative filtering algorithm, and the recommendation rate of the traditional algorithm is improved to a certain extent.

From Figure 2, the improved average absolute deviation MAE based on the user-interest matrix into the clustering idea is improved obviously under different neighborhoods. It is shown that the proposed algorithm is more effective than the traditional cooperative filtering algorithm, which further shows the effectiveness of the improved algorithm proposed in this paper when recommending the result to the target user.
Conclusions
This paper analyzes the traditional collaborative filtering algorithm ideas and algorithm steps. In this paper, a new algorithm is proposed based on the user-interest matrix integration and clustering based on the cooperative filtering recommendation, which is based on the problem of low real-time performance and low recommendation accuracy. The experimental results show that the new algorithm has a better idea than the traditional cooperative filtering algorithm in terms of real-time recommendation and recommendation accuracy, which provides a new idea for the future processing of large-scale data.

References