Algorithm Optimization for Cold Start of Collaborative Filtering System

Shuo Wang*, Jing Yang*, Qingchao Zhaob
Harbin Engineering University, China

*Corresponding author e-mail: yangjing@hebeu.edu.cn, *tuzi@hebeu.edu.cn,
bzhaoqc418@hrbeu.edu.cn

Abstract. With the development of network technology, the amount of information on the network has grown and expanded rapidly with an exponential law, and its information organization is heterogeneous, diverse, and distributed. With the vigorous development of Internet information services, the scale of its information resources has also exploded. For ordinary users, the problems of “information trek” and “information overload” on the Internet are becoming increasingly serious. In order to solve the problem of information overload, recommendation system has become an indispensable tool for today’s e-commerce platform, which can help users find valuable information quickly. Collaborative filtering-based recommendation algorithms have been widely applied and studied in recommendation systems. Although the collaborative filtering algorithm has been widely used, there are still problems such as data sparsity, scalability, and cold start, which seriously limit the quality of recommendations. Therefore, collaborative filtering algorithms face many challenges. Especially recommendation systems using collaborative filtering technology. This article discusses the cold start system, the cold start user, and the cold start scenario. Effective use of project content information and user personal information is one of the effective methods to solve the cold start problem, that is, a hybrid recommendation technology combining information filtering and collaborative filtering. This paper proposes a hybrid recommendation technology that can better solve the cold start problem.

1. Introduction
In the era of information overload, recommendation technology can help users effectively filter information. Collaborative filtering algorithms are widely used in recommendation systems due to their high efficiency and accuracy [1-3]. However, with the rapid increase in the number of projects and users in the recommendation system. The algorithm faces severe cold start problems, which greatly reduces the recommendation effect of the algorithm itself. Matrix decomposition model is one of the widely used recommendation algorithms in collaborative filtering systems. A large number of studies have shown that it is significantly better than other collaborative filtering algorithms in terms of recommendation speed and accuracy [4-6]. Collaborative filtering recommendation based on matrix factorization relies on the feature matrix of learning users and recommendation items for recommendation [7-9]. If any user has not rated a new item in the rating matrix. Or if the new user has not evaluated any items in the scoring matrix, he will not be able to learn the feature matrix of new users.
and new items [10]. In this way, new users and projects cannot generate recommendations for the application of their matrix factorization models. This is a common cold start problem in collaborative filtering systems [11, 12].

Currently, the most widely used collaborative filtering recommendation algorithm only considers historical score factors, and ignores user characteristics and item characteristics. Because in the historical scoring matrix, the number of users is usually much smaller than the number of items, which leads to sparse scoring data [13,14]. At the same time, collaborative filtering recommendation algorithms need to make recommendations based on historical rating data, because new users and new items do not have historical rating information [15]. Therefore, the recommendation system based on this algorithm cannot recommend to new users and new projects, resulting in a cold start problem. Traditional content-based recommendation algorithms make recommendations based on item characteristics only, without considering historical scoring factors [16]. However, due to the limited features of users and projects, and the difficulty of obtaining a large amount of valid data for the recommendation system, the accuracy of this recommendation algorithm is very low and cannot be effectively recommended to users.

In the collaborative filtering system, in order to increase the chance of being recommended for new projects and improve the recommendation effect for new users, the cold start problem should be effectively solved. In a general collaborative filtering system, there will be simple attribute information about users and projects. Other ways to solve the cold start problem include recommending the most popular recommendations for new users or recommending new items to users with broad interests. This article considers three cases of cold start system, cold start user and cold start project. Effective use of project content information and user personal information is one of the effective methods to solve the cold start problem, that is, a hybrid recommendation technology combining information filtering and collaborative filtering. This paper proposes a hybrid recommendation technique that can better solve the cold start problem.

2. Method

2.1. Research on Cold Start Problems

The cold start problem can be simply described as the lack of new user or new item scoring information, which will cause a series of problems, such as similarity calculation, nearest neighbor query and scoring prediction. Therefore, the breakthrough point of the problem can be started from three aspects, namely, filling in the blank of the score, modifying the similarity calculation formula, and replacing the similarity calculation with other methods. When the review data is sufficient, the collaborative filtering algorithm can produce effective recommendations. However, when the scoring matrix is extremely sparse, the collaborative filtering algorithm cannot provide high-quality recommendations. Some linear factor models have been proposed to solve the sparsity problem. Use methods such as singular value decomposition principal component analysis or maximum edge matrix factorization to reduce the dimensionality and smoothing noise of the user term matrix. However, these linear decomposition methods cannot solve the cold start problem. Some hybrid methods combining information filtering and collaborative filtering are used to solve the cold start problem. The cold start problem is a very serious problem for collaborative filtering algorithms. This problem can be divided into two categories: user cold start and project cold start. First, let's introduce the user's cold start. The problem is that when a new user uses the recommendation system for the first time, the user has not yet rated all the products in the system. Generate personalized recommendations for this user. A user cold start indicates that the project has a user rating of 1 to 5. As shown in Table 1 below, the empty score defaults to 0.
Table 1. User ratings

| User   | Project A | Project B | Project C | Project D |
|--------|-----------|-----------|-----------|-----------|
| User 1 | 3         | 4         | 3         | 4         |
| User 2 | 4         | 4         | 5         | 2         |
| User 3 | 3         | 3         | 5         | 5         |
| new user | 0       | 0         | 0         | 0         |

Whether it is a cold start of a user or a cold start of a project, the impact of this issue is a lack of scoring information. The lack of rating information prevents target users from having common rating items with any users in the system, so ratings are similar. That is, the denominator of the Pearson correlation coefficient is always zero, and the similarity calculation formula cannot be used normally. Because the similarity cannot be calculated, the system cannot provide users with personalized recommendations.

2.2. Basic Model of Collaborative Filtering Recommendation System

The goal of the collaborative filtering recommendation system is to predict the current user's score on un-scored items and give a list of recommendations. The recommendation list is usually divided into three parts: original data set, collaborative filtering predictor and candidate recommendation set. The original data set includes the original scoring matrix, user set, and item set. The predictor uses collaborative filtering algorithms to predict scores. The candidate recommendation set stores items of the prediction score. Collaborative filtering recommendation is a personalized recommendation technology often used in recommendation systems. It includes three types of collaboration, namely user-based, project-based, and model-based collaborative filtering. The following mainly uses user-based recommendation methods. User-based collaborative filtering algorithms mainly include building a user model, finding the nearest neighbors, and generating recommendations. In order to build a user model, user interest preferences need to be collected first, and display levels and implicit ratings are methods to obtain user interest preference information. The score reflects the relationship between the user and the product, and indicates the degree of user preference for the product. The display level is the user's direct rating of the product. There are multiple ratings, which can be explicit values or binary representations. Implicit scoring is the guessing of the user's interest by the recommendation system. It is an analysis of the user's behavior by the recommendation system, not the user's preference for the product. After obtaining the user interest preference information, we need to process this information to obtain the user item score matrix $R_{m*n}$, where the mth row of the matrix represents the number of users, the nth column of the matrix represents the number of items, and $R_{ij}$ represents me as the item j-rated users.

This algorithm is widely used because of its simplicity and reliability. However, this algorithm has the following two problems. The first problem is that it is sensitive to the choice of initial cluster centers. The second problem is that it cannot be determined in advance. This paper determines the upper limit of the $A$: value and verifies the rationality of formula (1).

$$k_{max} \leq \sqrt{n}$$ (1)

Therefore, the focus of this article is how to optimize the selection of the initial clustering center. Before introducing the optimization method, we first understand two concepts. At any point x in the data set $U$. Calculate the similarity between x and other points in the data set $D(x)$. The number of points whose similarity exceeds the threshold is used as the point density of the point, and the point density is expressed as follows:

$$D(x) = \left| \left\{ p | \text{sim}(x, p) \geq \varepsilon, p \in U \right\} \right|$$ (2)
Collaborative filtering technology is one of the most successful recommendation technologies currently in use. Its biggest advantage is that it does not need to analyze the characteristic attributes of the project, has no special requirements for the recommendation system, and can handle unstructured responsible projects. Usually, e-commerce recommendation systems are usually developed in a challenging environment, especially for large online shopping sites (such as Amazon, Taobao, etc.). Usually fast and accurate recommendation systems will arouse user interest and bring benefits to the company. For collaborative filtering recommendation systems, effectively addressing the following issues can lead to high-quality prediction scores or recommendations.

3. Experiment
The data set of this experiment is a public data set of 100,000 rating records provided by Movielens website, which includes 1943 users’ ratings on 1682 hotspots, each of which has been rated at least 10 words or more (rating value 1 to 5). The higher the value, the more interest the user has in the movie. The Movielens public dataset is a very common dataset in the field of personalized recommendations. The data set comes from 50-day behavior data of about 200 million registered users, including about 2 million active users, 6000 projects, and 300 million historical behavior records, and also includes social networks, user tags, project classification, rich contextual information, etc. Project keywords. However, such a huge data set is far beyond the computing power of the laboratory, so we extracted a smaller data set from it. The basic statistical information is shown in Table 2.

| Description        | User     |
|--------------------|----------|
| Article            | 11576    |
| Score              | 359465   |
| User Tag           | 13254    |
| Item Keywords      | 12022    |
| User Social Relation | 3214    |

According to the ratio of 80% and 20%, the selected data set is divided into a training set and a test set. For the cold start user problem, this paper selects users who have published at least 40 levels from the training data set. And these users evaluated at least one item in the test data set. The experimental data set selected 1241 cold-start users from it. These users evaluated at least one item in the test set and posted at least 40 reviews in the training set. After screening cold-start users, that is, the prediction quality of the prediction algorithm of new users, randomly divide these cold-start users into equal 5 different user sets, and choose to score these users respectively to form 5 different test set data. After this process, new users accounted for 84.13% of all users.

4. Discuss
Although the above work has improved the prediction accuracy of the traditional collaborative filtering algorithm to some extent, there are still problems such as data sparsity, cold start, and scalability, especially for the first two methods, whether the context filtering is performed first or later Execution context. When filtering, the filtered data may still be sparse, and still cannot get rid of the problem of sparse data. Although the context modeling method can find some potential relationships from the multidimensional sparse data model and alleviate the problem of data sparseness, the number of parameters in the model increases exponentially with the increase of data size and context information. From a runtime perspective, in cold start scenarios, cold start users, and cold start project scenarios. The feature-based bilinear regression model (FRBE), global averaging algorithm (AVG), and the most popular recommendation algorithm (MP) with bias are all better than the memory-based recommendation algorithm (filterbot algorithm). Figure 1 is the average time required for each algorithm to predict an item's score. Since many model-based collaborative filtering algorithms use simple matrix
product methods for prediction, memory-based recommendation algorithms usually spend more time looking for the most similar neighbors or entries.

A pattern is the most frequently occurring number in a set of data. It is the embodiment of centralized quantification. Its main idea is to use the pattern value of all user-rated items as the predicted score for unrated items. For example, since the user’s rating of the item is usually centralized, the user may score 4 points for the favorite item and 1 point for the less interesting item. Therefore, this model can be used to predict the score of new items for the current user. All current users’ rating patterns are used as the predicted scores for new items. Especially in a collaborative filtering recommendation system using a scoring threshold, the predicted value of a new item’s score may be lower than the scoring threshold, so it cannot be recommended. Figure 2 is the prediction results of various algorithms on a data set containing 10,000 score data.

---

**Figure 1.** Average run time of each algorithm run

**Figure 2.** Data changes as the training set grows
From Figure 2, we can see that the Bilinear Regression Model (FRBE), the Mean Value Algorithm (AVG), and the Most Popular Recommendation Algorithm (MP) with biased Gideon features perform much better than other algorithms. Bias-based bilinear regression model with biased features has the same effect when constructing project attributes, regardless of whether dynamic attributes are considered. It can be seen from Figure 2 that as the dataset sparsity continues to increase. The prediction effect of the most popular algorithm (MP) is hardly affected by the sparsity of the data system. The filterbot algorithm's prediction efficiency is not as good as other algorithms.

5. Conclusion
Collaborative filtering algorithm is one of the most successful methods to solve the problem of information overload. It records and analyzes the user's historical behavior, predicts the most likely behavior of future users, and better meets the personalized needs of users. However, with the rapid growth of the number of products and users, the problems of sparse data and cold start have gradually emerged. The problem with data sparseness is that the number of user ratings for a product is much smaller than the total number of products in the system, which will result in many blank entries in the rating matrix. Finally, the effectiveness of the cold-start algorithm proposed in this paper is verified on the actual data set. The experimental results show that compared with the algorithm proposed by the predecessors to solve the cold start problem of collaborative filtering systems. The algorithm proposed in this paper can better solve the cold start problem of new users and new projects in the collaborative filtering algorithm based on matrix factorization. The current social network is booming, and the application of Graph neural network (GNN) in social network has become a hot topic. In the future research, we will also study how to use GNN to solve the cold start problem and relationship reasoning in recommendation system diagram.

References
[1] Wei J, He J, Chen K, et al. Collaborative filtering and deep learning based recommendation system for cold start items[J]. Expert Systems with Applications, 2017, 69:29-39.
[2] Yu C H, Liu X J, Li B, et al. Mobile Service Recommendation Based on Context Similarity and Social Network[J]. Tien Tzu Hsueh Pao/Acta Electronica Sinica, 2017, 45(6):1530-1536.
[3] Zhang X, Zhao H. Cold-Start Recommendation Based on Speech Personality Traits[J]. Journal of Computational and Theoretical Nanoscience, 2017, 14(3):1314-1323.
[4] Mansoury M, Shajari M. Improving recommender systems’ performance on cold-start users and controversial items by a new similarity model[J]. International Journal of Web Information Systems, 2016, 12(2):126-149.
[5] Wang X, Luo F, Qian Y, et al. A Personalized Electronic Movie Recommendation System Based on Support Vector Machine and Improved Particle Swarm Optimization[J]. PLoS ONE, 2016, 11(11): e0165868.
[6] Afify Y M, Moawad I F, Badr N L, et al. A personalized recommender system for SaaS services: A Personalized Recommender System for SaaS Services[J]. Concurrency and Computation Practice and Experience, 2016, 29(4).
[7] Xu J, Yao Y, Tong H, et al. RaPare: A Generic Strategy for Cold-Start Rating Prediction Problem[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, PP (99):1-1.
[8] Cui Q, Wu S, Liu Q, et al. MV-RNN: A Multi-View Recurrent Neural Network for Sequential Recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, PP (99):1-1.
[9] Diego Sánchez-Moreno, Ana B. Gil González, M. Dolores Muñoz Vicente, et al. A collaborative filtering method for music recommendation using playing coefficients for artists and users[J]. Expert Systems with Applications, 2016, 66(C):234-244.
[10] Vinayak S, Sharma R, Singh R. MOVBOK: A Personalized Social Network Based Cross Domain Recommender System[J]. Indian Journal of Science and Technology, 2016, 9(31).
[11] Yang B, Lei Y, Liu J, et al. Social Collaborative Filtering by Trust[J]. IEEE Transactions on
Pattern Analysis and Machine Intelligence, 2017, 39(8):1633-1647.

[12] Li J., Yang J J., Zhao Y., et al. Enforcing Differential Privacy for Shared Collaborative Filtering[J]. IEEE Access, 2016, 5:1-1.

[13] Kharrat F B., Elkhleifi A., Faiz R., et al. Improving Collaborative Filtering Algorithms: Sentiment-based Approach in Social Network[J]. International Journal of Knowledge Society Research, 2016, 7(3):99-118.

[14] Liu C L., Wu X W. Large-scale Recommender System with Compact Latent Factor Model[J]. Expert Systems with Applications, 2016, 64:467-475.

[15] Miao Z., Yan J., Chen K., et al. Joint Prediction of Rating and Popularity for Cold-Start Item by Sentinel User Selection[J]. IEEE Access, 2016, PP(99):1-1.

[16] Gohari F S., Haghighi H., Aliee F S. A semantic-enhanced trust based recommender system using ant colony optimization[J]. Applied Intelligence, 2016, 46(2).