Optimization of Collaborative Filtering Algorithm in Movie Recommendation System

Jiao Peng(✉) and Shu Gong

Guangdong University of Science and Technology, Dongguan, Guangdong, China
154711901@qq.com

Abstract. In the era of big data information explosion, people are faced with a large amount of information every day, and how to obtain the required content in a large amount of information. The appearance of intelligent recommendation system has brought great convenience to our life. The recommendation system can recommend corresponding functions, products and services according to users' past browsing information, enabling users to get their desired information data from massive data more efficiently. As an indispensable part of most people's entertainment life, movie recommendation has also become a very important part of Internet recommendation content. The collaborative filtering algorithm is used to realize the personalized recommendation of movies. However, in the process of movie recommendation, it is found that new users are only recommended movies based on their ratings without considering the attribute information between movies, which may lead to problems such as inaccurate recommendation accuracy. Therefore, this paper further optimizes the collaborative filtering algorithm and introduces the similarity calculation between movie attributes to improve the accuracy of movie recommendation.

Keywords: Movie recommendation system · Collaborative filtering algorithm · Optimization of algorithm

1 Introduction

In today's information explosion, there are various ways and means of obtaining information, as well as various kinds of information. Now what people spend the most time is no longer where to get information, but to find the content or information they are interested in among the numerous information, which is the so-called information overload problem. To solve this problem, recommendation systems are applied. Recommendation systems are ubiquitous in everyday web applications, such as online shopping, online bookstores, news apps, social networks, music websites, movie websites and so on. The system will make personalized content recommendations based on personal preferences, habits or needs. For example, open the news app. Because of the personalized recommendation function in the app, the front page of news is different for everyone. Collaborative filtering algorithm is the first personalization recommendation technology and the most widely used recommendation algorithm. The movie recommendation system makes use

X. Chen et al. (Eds.): ML4CS 2020, LNCS 12488, pp. 11–19, 2020.
https://doi.org/10.1007/978-3-030-62463-7_2

of this special information recommendation technology to recommend some movies that users may like or be interested in according to their previous information about watching movies and their rating information. However, in the process of using collaborative filtering algorithm, we found that if the movie recommendation is only based on the user's rating of the movie, the recommendation result is not accurate, and the recommendation accuracy and coverage may not be accurate enough. Therefore, collaborative filtering algorithm can be further optimized. In the process of movie recommendation, not only the rating information of the movie by users, but also the attribute information of the movie itself, such as the movie category and the similarity between the movie categories.

2 Collaborative Filtering Algorithm

2.1 Introduction of Collaborative Filtering Algorithm

Collaborative filtering algorithm is a recommended technique which is widely used and matures at present. Its working principle is to learn the user's interests and preferences according to the user's historical operation behavior, find out the user's neighbor users according to their interests and hobbies, determine the set of goods that the neighbor user likes, and then recommend to the user the information of goods in the set that the user has not bought.

The collaborative filtering algorithm is mainly divided into three steps:

(1) Create a "user-item" rating table

According to the user's purchase record and the score information of the purchased items, a "user-item" rating table is formed. Among them, The user set $U = \{u_1, u_2, \ldots, u_m\}$, where u_i represents the user i, $T = \{t_1, t_2, \ldots, t_n\}$, where t_i represents the item t. R_{u_i,t_i} represents user u_i buys item t_i and gives a rating. The "user-item" rating table formed is shown in Table 1:

Table 1. "user-item" rating table

	t_1	t_2	…	t_n
u_1	R_{u_1,t_1}	R_{u_1,t_2}	…	R_{u_1,t_n}
u_2	R_{u_2,t_2}	R_{u_2,t_2}	…	R_{u_2,t_n}
…	…	…	…	…
u_m	R_{u_m,t_1}	R_{u_m,t_2}	…	R_{u_m,t_n}

(2) Form a neighborhood set

Neighbor set refers to the collection of users (items) with common characteristics with the target user. According to the data set in Table 1, the historical score records of each user (item) can be obtained from the "user-item" scoring matrix. Assuming that the target user is u_v, to calculate the set of neighbors of the user u_v, you should first

calculate the similarity between the target user and other users. The higher the similarity between users, the closer the two users' interests and hobbies will be, and the greater the probability that they will share the same characteristics. Sort by similarity degree, take the first N users to form the neighbor set.

(3) Form a list of recommendations

Set the user's neighborhood set $U = \{u_1, u_2, \ldots, u_N\}$, and then add the collection of favorite items of each neighbor user to the list to form a recommended list set, which is recommended to the user.

Collaborative filtering recommendation algorithm can also be subdivided into two algorithms: user-based collaborative filtering and item-based collaborative filtering recommendation algorithm.

The main principle of user-based collaborative filtering algorithm: When the user needs personalized recommendation, according to the user's personal interests, hobbies or behavior habits and other information, calculate the similarity between users, and find the set of neighbors with the user. Then the information that each user likes in the neighbor set is added together to form a commodity set, and then the items in the commodity set are recommended to the user.

The main principle of item-based collaborative filtering algorithm: When a user needs personalized recommendation, he can usually analyze the historical behavior data of the goods he has purchased before, analyze the characteristics of the goods he has purchased, and then learn the characteristics of the items he likes. In the commodity set, according to the user's interest and the characteristics of the items to be recommended, a group of items with the greatest correlation is determined in the commodity set as the recommendation list and recommended to the user.

2.2 The Evaluation Index of Recommendation Algorithm

The evaluation index is used to measure the performance of collaborative filtering recommendation algorithm. The recommendation performance is mainly reflected in two aspects: prediction quality and recommendation quality. Recommendation algorithms usually use precision rate and recall rate to measure the quality of recommendation, and use coverage rate to measure the quality of recommendation prediction by recommendation system.

(1) Precision rate

The precision rate is the measure to evaluate the recommendation algorithm, which is used to evaluate the recommendation effect of the recommendation system. Precision rate is used to describe the ability of a recommendation system to predict user behavior. Generally, the coincidence rate between the recommendation list and user behavior given by the algorithm on the offline data set is calculated. The higher the coincidence rate is, the higher the accuracy will be.

$$\text{Precision} = \frac{\sum_{i=1}^{m} \left| T_{u_i} \cap recommender_{u_i} \right|}{\sum_{i=1}^{m} \left| T_{u_i} \right|} \tag{1}$$

Where, Precision refers to the precision of the recommendation algorithm, T_{u_i} refers to the items purchased by user u_i, and $recommender_{u_i}$ refers to the items recommended to user by the system. $T_{u_i} \cap recommender_{u_i}$ refers to the items that user u_i purchases which are recommended to the user by the recommendation system.

(2) Recall rate

Recall rate is a measure of how good a recommendation algorithm is. Recall rate measures the quality of the recommendation system by calculating the proportion of items purchased in the recommended items.

$$\text{Recall} = \frac{\sum_{i=1}^{m} \left| T_{u_i} \cap recommender_{u_i} \right|}{\sum_{i=1}^{m} \left| recommender_{u_i} \right|} \tag{2}$$

Where, Precision refers to the precision of the recommendation algorithm, T_{u_i} refers to the items purchased by user u_i, and $recommender_{u_i}$ refers to the items recommended to user by the system. $T_{u_i} \cap recommender_{u_i}$ refers to the items that user u_i purchases which are recommended to the user by the recommendation system.

(3) Coverage rate

Coverage rate is an important measure of a recommendation system. Coverage is calculated as the ratio of the number of items recommended to the user to the total number of items. Suppose the user set of the system is U, the list of items is represented by L, and the length of the list of items recommended to users is represented by R(u).

$$\text{Coverage} = \frac{\left| \sum_{u \in U} R(u) \right|}{|L|} \tag{3}$$

3 Optimization of User-Based Collaborative Filtering Algorithm

3.1 User-Based Collaborative Filtering Algorithm

The main content of user-based collaborative filtering recommendation algorithm is: when to recommend items to target users A, with target users will find A have similar preferences, user B, then the user B items like, the target users before A no goods, recommended to the user. For example, in Fig. 1, we can see that user u_1 watched the movie m_1 and m_2, user u_2 watched the movie m_1, m_3 and m_4, user u_3 watched the movie m_1, m_5 and m_6. Recommended new movies to user u_1 based on the above viewing records.

The relationship between movie names and movie types is shown in Table 2.

(1) According to the similarity between users, find the set of users with similar preferences to the target users.

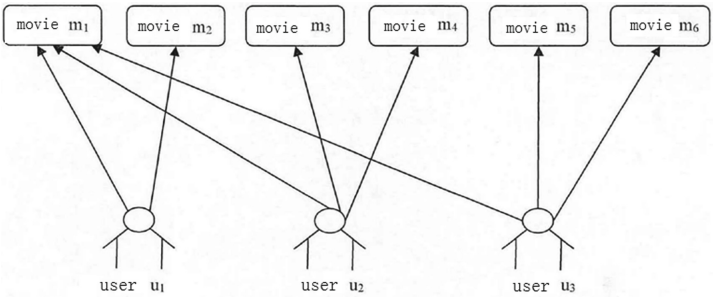

Fig. 1. User-based collaborative filtering recommendation

Table 2. Relationship between movie titles and movie types

Name/type	Romance	Comedy	Drama	Action	Horror	Thriller	Fantasy	Crime
m_1	1	1	1	0	0	0	0	0
m_2	1	0	0	1	0	0	0	0
m_3	1	0	0	0	1	0	0	0
m_4	1	0	0	0	0	1	0	0
m_5	1	0	0	0	0	0	1	1
m_6	1	0	0	0	0	0	1	1

For the first time, we need to calculate the similarity between users. We set user u_i and u_j, $\mathbf{N}u_i$ represents the collection of items that user u_i likes, and $\mathbf{N}u_j$ represents the collection of items that user u_j likes. We can use the following formula to calculate the similarity between users.

$$\text{sim}\,(u_i, u_j) = \frac{\left|\mathbf{N}u_i \cap \mathbf{N}u_j\right|}{\sqrt{|\mathbf{N}u_i|\left|\mathbf{N}u_j\right|}} \tag{4}$$

In Fig. 1, user u_1 likes to watch movie m_1, m_2, user u_2 likes to watch movie m_1, m_3 and m_4, user u_3 likes to watch movies m_1, m_5 and m_6. Now recommend movies to target user u_1. Calculate the user similarity between user u_1 and user u_2.

$$\text{sim}(u_1, u_2) = \frac{|\{m_1, m_2\} \cap \{m_1, m_3, m_4\}|}{\sqrt{|\{m_1, m_2\}||\{m_1, m_3, m_4\}|}} = \frac{|\{m_1\}|}{\sqrt{6}} = \frac{1}{\sqrt{6}}$$

The user similarity between user u_1 and user u_3 is calculated as follows.

$$\text{sim}(u_1, u_3) = \frac{|\{m_1, m_2\} \cap \{m_1, m_5, m_6\}|}{\sqrt{|\{m_1, m_2\}||\{m_1, m_5, m_6\}|}} = \frac{|\{m_1\}|}{\sqrt{6}} = \frac{1}{\sqrt{6}}$$

The similarity between user u_1 and user u_2 is calculated to be the same as that between user u_1 and user u_3.

(2) Forming neighbor sets and recommendation lists

By the above calculation, $\{u_2, u_3\}$ becomes the user set of user u_1. Add user u_2 and u_3's movies to the list to form a recommendation list $\{m_3 , m_4 , m_5 , m_6\}$,which is recommended to user u_1.

However, from Table 2, we can see that user u_1 likes movie m_1 and m_2, while movie m_1 and m_2 belong to romance, so user u_1 prefers Romance. User u_2 likes to watch movie m_1, m_3, and m_4, while movie m_1, m_3, and m_4 belong to romance, so user u_2 prefers romance. User u_3 likes to watch movie m_1, m_5, and m_6, while movie m_5, and m_6 belong to science fiction and crime movies, so user u_3 prefers science fiction and crime movies. To sum up, user u_1 and user u_2 prefer romance movies, while user u_3 prefers science fiction and crime movies. Therefore, user u_1 has a high similarity with user u_2, but a low similarity with user u_3.

Based on the above analysis, it can be seen that when using user-based collaborative filtering algorithm for movie recommendation, the attribute characteristics of items that users like need to be considered; otherwise, the similarity between users calculated is incorrect. Therefore, it is necessary to optimize the user-based collaborative filtering algorithm.

3.2 Optimization of User-Based Collaborative Filtering Algorithm

When the collaborative filtering algorithm based on users is optimized, the attribute information of items is added in the process of similarity calculation. The specific steps are as follows: firstly, the collaborative filtering algorithm is used to calculate the similarity between users, then the similarity of attribute feature vectors of the items that users like is calculated, and finally the two similarities are fused as the final similarity between users. Make recommendations based on the final similarity calculation.

(1) The first step is to calculate the similarity between users.

$$\text{sim}\,(u_i, u_j) = \frac{\left|\mathbf{N}u_i \cap \mathbf{N}u_j\right|}{\sqrt{\left|\mathbf{N}u_i\right|\left|\mathbf{N}u_j\right|}}$$

(2) Calculate the similarity of properties between films

Set the category set of the movie $A = \{a_1, a_2, \ldots, a_n\}$, and construct the movie's attribute feature vector according to the category to which the movie belongs. $\underset{a_{m_i}}{\rightarrow} = \{a_{i1}, a_{i2}, \ldots, a_{in}\}$, $\underset{a_{m_i}}{\rightarrow}$ represents film m_i category information vector, as shown in Fig. 2, a_{ij} value is 0 or 1, used to indicate whether movies m_i belongs to the category a_j. Calculate the similarity between films as shown in formula (5).

$$\text{sim}\left(m_i, m_j\right) = \underset{a_{m_i}}{\rightarrow} * \underset{a_{m_j}}{\rightarrow} = \frac{\sum_{b=1}^{S}\left(a_{ib} * a_{jb}\right)}{\sqrt{\sum_{b=1}^{S} a_{ib}^2}\sqrt{\sum_{b=1}^{S} a_{jb}^2}} \tag{5}$$

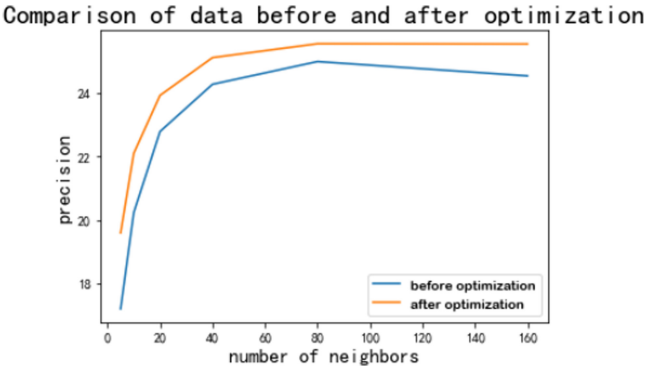

Fig. 2. Comparison of precision before and after optimization

In the formula, b represents the category b attribute of the movie; S represents the total S attributes of the movie, and $\text{sim}(m_i, m_j)$ represents the similarity between movie m_i and movie m_j.

(3) Calculate the similarity between users after optimization

The similarity between users after optimization mainly considers the similarity of item attributes while calculating the similarity of users. The specific calculation formula is shown in formula 6:

$$\text{sim}(u_i, u_j) = \alpha * sima + (1 - \alpha) * simb \tag{6}$$

In the formula, $\text{sim}(u_i, u_j)$ is the similarity between the optimized users, *sima* is the similarity of item attribute characteristics, *simb* is the user similarity calculated by the traditional collaborative filtering algorithm, and α is the parameter to adjust the two similarities.

(4) Forming neighbor sets and recommendation lists

According to the optimized user similarity value, users with higher similarity value constitute neighbor sets, and the recommendation list is constructed to complete the user's movie recommendation.

4 Algorithm Evaluation Experiment

4.1 Data Acquisition

Using Movie Lens data set, the collection contains 209,171 movies rated by 526 users with a score of 0–5. The movie categories are mainly divided into 12 categories, including adventure, animation, children, drama, fantasy, romance, drama, action, crime, thriller, horror and so on. Category information for movie properties is indicated by 0 and 1 respectively to indicate whether the movie belongs to that category. The data in the set is

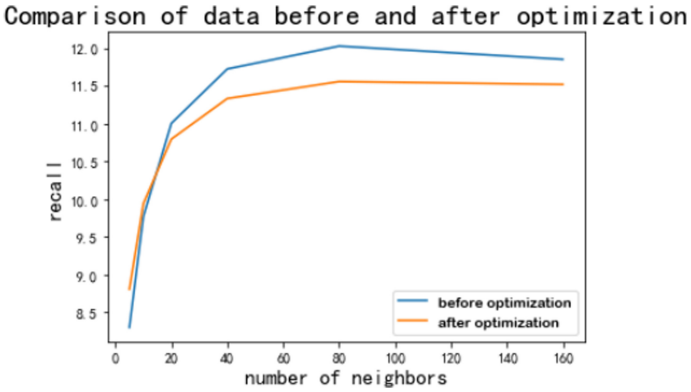

Fig. 3. Comparison of recall rate before and after optimization

divided into two parts, one as the training set and the other as the test set. Select different number of neighbors to test.

4.2 Algorithm Testing

Using the data collected above, the optimization algorithm based on user collaborative filtering was used to calculate the precision rate, recall rate and coverage rate before and after optimization. By comparing the three evaluation indexes before and after optimization, data simulation test was conducted according to formulas (1), (2) and (3). The specific test results are shown in Fig. 2, Fig. 3 and Fig. 4.

As can be seen from the following three figures, in the user-based collaborative filtering algorithm, the similarity of attributes and features of items is introduced to calculate the user similarity. After the algorithm is optimized, the precision and coverage of movie recommendation are significantly improved, but at the cost of the recall rate.

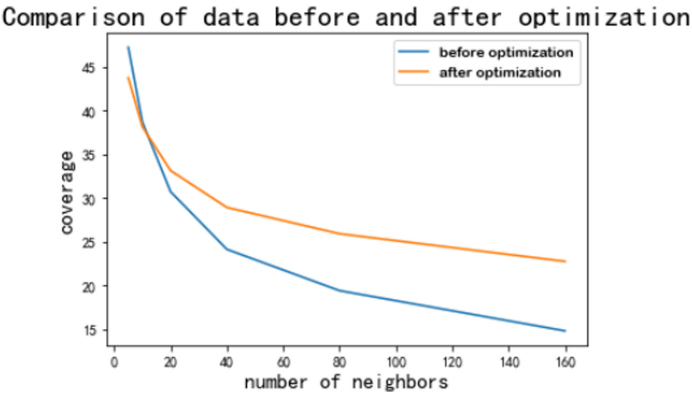

Fig. 4. Comparison of coverage before and after optimization

5 Conclusion

This article mainly adopted the user-based collaborative filtering algorithm to movie recommendation, but in the process of recommendation, find only according to the similarity between the user and the target collection of users, the user preference similarity recommendation precision is not accurate, so the film was introduced to the optimization algorithm in attribute characteristics of similarity between two neighbor set and form the suggestion list is calculated after synthetically, in order to enhance the precision in movie recommendation.

References

1. Chen, J.F., Huang, L.S., Lin, G.T.: TV program recommendation strategy of collaborative filtering algorithm. J. Henan Inst. Eng. (Nat. Sci.) **32**(01), 61–65 (2020)
2. Jiao, J.F.: Research on collaborative filtering recommendation algorithm based on Spark. Comput. Program. Skills Maintenance **03**, 40–41 (2020)
3. Li, Z.F., Liu, Y.S., Li, C.T.: Research on news recommendation algorithm based on user behavior. Comput. Eng. Sci. **42**(03), 529–534 (2020)
4. Dong, Y.F., Zhu, C.S.: Collaborative filtering algorithm based on improved user attribute rating. Comput. Eng. Des. **41**(02), 425–431 (2020)
5. Ding, H.F., Huang, Q.S.: Research on personalized tourism recommendation algorithm based on attribute characteristics. Intell. Comput. Appl. **10**(01), 193–196 (2020)