Improvement and Research of Collaborative Filtering Algorithm Based on Penalty Factor

Panying Li¹, Yuxuan Han¹, Xiumei Wen²* and Fanxing Meng¹*

¹ Computer Science and Technology, Hebei Institute of Civil Engineering and Architecture, Zhangjiakou City, Hebei Province, Zip Code: 075000, China
² Zhangjiakou Big Data Technology Innovation Center, Zhangjiakou City, Hebei Province, Zip Code: 075000, China
*Corresponding author’s e-mail: xiumeiwen@163.com.cn, sir363@163.com

Abstract-The rapid development of the Internet has pushed society into the era of information explosion, and people are faced with more and more information screening and choices. A recommendation system is an effective way to process massive amounts of information, and it is also a tool that can make recommendations based on user behavior. Traditional collaborative filtering algorithms generally use the cosine similarity formula to calculate the similarity between users or items to make recommendations. Due to the popularity of the Internet, more and more popular items have appeared. The appearance of popular items not only affects the recommendation results, but also fails to reflect the real needs of users. This paper proposes an improved formula for cosine similarity with a penalty factor, which can restrain the influence of popular items on the recommendation result. Finally, the Movie Lens data set is used to verify that the recommended performance indicators have been improved to a certain extent.

1. Introduction
In 1994, researchers from the Group Lens project research group at the University of Minnesota developed the first Group Lens system based on collaborative filtering. The basic idea of this system design is to analyze and mine users who have the same interests and preferences as the target user through the user's evaluation of the project, and finally give a recommendation list. The essence of the system is to use a user-based collaborative filtering algorithm. Collaborative filtering algorithm has many advantages in application. This technology is easy to implement, easy to operate, and has a good recommendation effect. Therefore, this technology has made great progress in both academic research and practical application. Domestic research on recommender systems is not very early, but some domestic researchers have also achieved certain results in recent years. At present, it mainly focuses on collaborative filtering algorithm and hybrid recommendation algorithm. In terms of commercial applications, many domestic companies have participated in research related to recommender systems. Among domestic companies, Percentage is the first to research and recommend technology. In order to quickly promote recommendation technology in China, the company has carried out a lot of technology sharing, provided personalized solutions, and developed China's recommendation technology. At the same time, domestic Internet company giant Alibaba also uses many recommendation techniques on Taobao and other websites. Therefore, recommendation systems have important applications in all walks of life. Many algorithms have emerged in the field of recommendation systems. These algorithms
have their own application scenarios, but at the same time, these algorithms also have certain advantages and disadvantages. Commonly used recommendation algorithms mainly include content-based recommendation, collaborative filtering recommendation and model-based recommendation algorithm, etc. Collaborative filtering recommendation algorithms are further divided into user-based collaborative filtering and item-based collaborative filtering. Among them, the most commonly used recommendation algorithm is the collaborative filtering recommendation algorithm. The operation steps of the algorithm are simple and clear, easy to understand, and have strong applicability. However, the collaborative filtering algorithm still has some problems, such as cold start, sparse data, poor scalability, etc. These shortcomings will affect the performance of the recommendation system in practical applications.

This article starts with the shortcomings of the collaborative filtering recommendation algorithm. Aiming at the traditional collaborative filtering recommendation algorithm that uses the cosine similarity calculation formula for recommendation, the impact of popular items cannot be suppressed, so the idea of adding a penalty factor to the cosine similarity formula is proposed. Finally, the MovieLens data set is used for experimental verification and comparative analysis of the results. The idea and implementation steps of user-based and item-based collaborative filtering recommendation algorithms are relatively similar. Therefore, the following uses user-based collaborative filtering recommendation algorithm as an example for experimental analysis and demonstration.

2. Overview of collaborative filtering recommendation algorithm

2.1 Traditional collaborative filtering algorithm
The user-based collaborative filtering recommendation algorithm generally needs to obtain the user's behavior information, and then find the neighbor set of the target user, and finally select the items in the neighbor set that are of interest to the target user but have not generated behaviors for recommendation. The specific implementation steps are as follows:

2.1.1 Collect user behavior data and generate user-item rating matrix
The simplest form of user behavior data on the website is the log[1], of course, there will be other forms as well. For these behavior data, preprocessing is required first. Preprocessing is to convert the user's historical behavior described in natural language into digital information [2]. After processing, the user's behavior information on the item can be transformed into a matrix R(m, n), which is called a user-scoring matrix.

2.1.2 Establish a set of neighbors with similar interests of target users
The key to this step is to calculate the interest similarity between the target user and other users, and generate the target user's neighbor set according to the user similarity. The similarity between users is calculated by the similarity formula. Common similarity calculation formulas include Pearson correlation coefficient, Euclidean similarity, and cosine similarity calculation formulas. The more commonly used is the cosine similarity calculation formula, the formula is as follows:

\[
W_{uv} = \frac{|N(u) \cap N(v)|}{\sqrt{|N(u)| \cdot |N(v)|}}
\]

(1)

Among them, N(u) represents a collection of items with positive feedback that user u has ever had, and N(v) represents a collection of items with positive feedback that user v has ever had. Positive feedback can also be called explicit feedback behavior, which includes the behavior of users expressly expressing their preferences for items [3].
2.1.3 Generate recommendations
After obtaining the interest similarity between users, the user-based collaborative filtering recommendation algorithm will recommend the most similar k favorite items to the user. The following formula measures the user u’s interest in items in the user-based collaborative filtering recommendation algorithm:

$$P(u,i) = \sum_{v \in S(u,k) \cap N(i)} W_{uv} \cdot R_{vi}$$

(2)

Among them, $S(u,k)$ contains k users whose interests are closest to user u, $N(i)$ is the set of users who have acted on item i, $W_{uv}$ is the similarity of interest between user u and user v, and $R_{vi}$ represents user v is interested in item i, because it uses implicit feedback data of a single behavior, and the value is 1 here. Implicit feedback data sets refer to behaviors that cannot clearly reflect user preferences, and the typical representative is user page browsing behavior [3].

2.2 Maintaining the Integrity of the Specifications
The algorithm flow of the recommendation system is generally divided into the following parts, namely the collection of user behavior data, the establishment of a user-item score matrix, the calculation of user similarity, and the generation of a user's neighbor set, and finally a recommendation list. The flowchart of collaborative filtering recommendation algorithm is shown in Figure 1.

![Collaborative filtering recommendation algorithm flow](image)

Figure1. Collaborative filtering recommendation algorithm flow

3. Research on collaborative filtering recommendation algorithm based on penalty factor

3.1 Penalty factor
Take books as an example. Two users have bought "Xinhua Dictionary". This does not mean that their interests are similar. If both users have bought "Introduction to Data Mining", then they can be considered to have similar interests, because Only people who study data mining will buy this book. Two users who have taken the same behavior on unpopular items can better illustrate the similarity of their interests.
In a recommendation system, if there are too many popular items, the recommendation results will be relatively single, without novelty, and the final recommendations are all items that appear on the popular rankings, which is very different from the actual needs of users. In order to reduce the impact of popular items, this paper considers adding a penalty factor as a weighting coefficient when calculating user similarity, which plays a role in restraining popular items from affecting the recommendation results. The more an item appears, the less the item contributes to the similarity of user interest. The revised formula can attenuate the impact of popular items. The revised formula is as follows:

\[ W_{uv} = \frac{1}{\sum_{i \in N(u) \cap N(v)} \log(1 + \sqrt{|N(i)|})} \]

\[ \sqrt{|N(u)| \cdot |N(v)|} \]

(3)

N(i) in the formula represents the number of occurrences of product i, and the impact of popular items on similarity is penalized by the number of occurrences of the item.

3.2 Improved algorithm flow
In the above process step of calculating user similarity, a penalty factor is added, and the penalty factor can be used to reduce the impact of popular items. The more items appear, the greater the effect of the penalty factor, and the smaller the impact of popular items on the calculation of user similarity, which has a good recommendation effect. The flow chart of the improved recommendation algorithm is shown in Figure 2.

4. Experimental results

4.1 Data set analysis
This article uses the classic Movie Lens (ml-100k) data set in the recommendation system to verify some performance indicators of the traditional user-based collaborative filtering recommendation algorithm and the improved user-based collaborative filtering recommendation algorithm. There are a total of three tables in the data set: Users, Movies and Ratings. The data set stores more than 100,000
pieces of user rating data for movies, of which the data rating value is 1 to 5. The data set is shown in Table 1.

| Category          | Quantity                                                                 |
|-------------------|---------------------------------------------------------------------------|
| Subscribers       | A total of 610, uid from 1 to 610                                         |
| Project volume    | A total of 1682, mid from 1 to 1682                                       |
| Score value       | 100k, an integer of 1-5                                                  |
| Sparsity          | 100k/(610*1682)=0.097                                                    |

When the system provides recommendation services, it generally provides users with a personalized recommendation list. This type of recommendation is called Top-N recommendation [4]. The prediction accuracy recommended by Top-N is generally measured by precision and recall.

R(u) is a recommendation list made to the user based on the user's behavior on the training set, and T(u) makes the user's behavior list on the test set. The accuracy rate describes how much of the final recommendation list is the user-item rating record that has occurred. The recall rate describes the percentage of user-item rating records included in the final recommendation list.

4.2 Experimental design
The offline experiment of the collaborative filtering algorithm is generally designed as follows: First, the user behavior data set is randomly divided into M parts according to a uniform distribution (M=8 in this experiment), one part is selected as the test set, and the remaining M-1 parts are used as the training set. Then build a user interest model on the training set, predict user behavior on the test set, and calculate the corresponding evaluation indicators. In order to ensure that the evaluation index is not the result of overfitting, M experiments are required, and a different test set is used each time. Finally, the average value of the evaluation indicators measured in M experiments is used as the final evaluation indicator.

4.3 Experimental results and analysis
In the verification experiment, this paper uses the classic cosine similarity algorithm and the improved cosine similarity calculation method for comparative analysis. First, the recommended number in the algorithm is set to 25, and the number of neighbors of the target user is 2, 3, 5, 10, 20, 40, 50. The calculation accuracy rate and recall rate (the calculation result retains four decimal places) are shown in Table 2.

| Number of similar users | Accuracy | Recall rate |
|-------------------------|----------|-------------|
| 2                       | 0.1622   | 0.0822      |
| 3                       | 0.1755   | 0.0888      |
| 5                       | 0.2063   | 0.1043      |
| 10                      | 0.2298   | 0.1162      |
| 20                      | 0.2500   | 0.1273      |
| 40                      | 0.2604   | 0.1302      |
| 50                      | 0.2615   | 0.1322      |

The improved cosine similarity algorithm is used to calculate the accuracy and recall rates. As shown in Table 3.
Table 3 Improved cosine similarity algorithm

| Number of similar users | Accuracy | Recall rate |
|-------------------------|----------|-------------|
| 2                       | 0.1670   | 0.0836      |
| 3                       | 0.1816   | 0.0915      |
| 5                       | 0.2083   | 0.1046      |
| 10                      | 0.2325   | 0.1171      |
| 20                      | 0.2566   | 0.1295      |
| 40                      | 0.2636   | 0.1326      |
| 50                      | 0.2662   | 0.1342      |

Set the input parameters (the number of user neighbors) to 2, 3, 5, 10, 20, 40, 50, and the recommended number to 25. The evaluation results are shown in Figure 3 and Figure 4.

Figure 3. Comparison of accuracy of traditional and improved recommendation algorithms

Figure 4. Comparison of the recall rate of traditional and improved recommendation algorithms
The figure above shows the comparison between the traditional collaborative filtering recommendation algorithm and the collaborative filtering recommendation algorithm with penalty factor added. It can be seen from the figure that the improved collaborative filtering recommendation algorithm has a higher accuracy and recall rate than the traditional collaborative filtering recommendation algorithm.

5. Conclusion
This article first analyzes that when the traditional collaborative filtering recommendation algorithm calculates user similarity, popular items will affect the final recommendation result. Using traditional cosine similarity algorithm for recommendation, the final recommended items are almost all popular items. Therefore, This paper proposes to add a penalty factor to the cosine similarity calculation formula to reduce the impact of popular items. Finally, it is verified by experiments that the improved algorithm can improve the accuracy and recall rate. It is concluded that the collaborative filtering algorithm based on the penalty factor proposed in this paper can improve the recommendation efficiency and make the recommendation result more accurate.

Acknowledgments
Thank you to the initiator of this research-”Hebei Provincial Higher Education Fundamental Scientific Research Business Expenses Research Project 2021XSTS04”. In addition, I would like to thank Zhangjiakou Big Data Technology Innovation Center for providing experimental platforms and equipment.

References
[1] Xiang,L.(2012) Recommended system practice.People Post Press,Beijing.
[2] Meng,XW. Hu,X. Wang,LC.(2013) Mobile recommendation system and its application.journal of Software .24(1):91-108.
[3] Meng,XW. Ji,WY. Zhang,YJ.(2015) Recommendation system in big data environment. Journal of Beijing University of Posts and Telecommunications. 38(2):1-15.
[4] Leng,YJ. Lu,Q. Liang,CY.(2014) Overview of Collaborative Filtering Recommendation Technology. Pattern recognition and artificial intelligence. 27(8):720-734.