2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA) | 978-1-6654-4276-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICPECA53709.2022.9719134

Research on Collaborative Filtering Recommendation Algorithm

Yuan Qingmin, Chen Xingyu
School of Management, Tianjin University of Technology
Tianjin, China
472457132@qq.com

***Abstract*—As a crucial technology in personalized recommendation, collaborative filtering algorithm has been widely used. Today's collaborative filtering algorithm is mainly divided into two types, user-based and commodity-based. In order to make it clearer to user-based with regard to the differences and application scenarios of the two collaborative filtering recommendation algorithms from the project, this article analyzes the definitions, basic principles, similarity calculation methods, and the advantages and disadvantages of the two algorithms through the method of comparison. Finally, through analysis and comparison, the current problems of these two algorithms and their respective limitations are proved to provide references for future use and improvement of collaborative filtering algorithms.**

Keywords—Collaborative filtering, recommendation algorithm, similarity

I. INTRODUCTION

Collaborative filtering recommendation algorithm is the most famous recommendation algorithm, and its main function is prediction and recommendation. [1] Coordinated filtering is the current mainstream type of recommendation algorithm, with a wide variety of tricks, and has been widely used in the industry. Its advantage is that it does not require much knowledge in a specific field, and better recommendation results can be obtained through statistical-based machine learning algorithms. The biggest advantage is that it is easy to implement in engineering and can be easily applied to products. So far, many recommendation algorithms have been proposed. Collaborative filtering is one of these algorithms. The most widely used and most effective recommendation algorithm. In modern commercial recommendation systems, most of them also apply collaborative filtering algorithms, and they have achieved certain success, because in fact, both user-based and item-based methods can achieve high recommendation accuracy. Although the collaborative filtering recommendation algorithm has been widely recognized in commercial systems, it does not mean that the applied algorithm is perfect. The collaborative filtering recommendation algorithm still has many unsolved problems, such as data sparseness and cold start problems. With the rise and application of the Internet, various social software and shopping software have followed. User-centric social networking sites have generated massive amounts of data related to user interests. So how can we use these data more effectively to seek recommendation algorithms? The promotion and improvement of the research has become the research topic of many scholars nowadays.

II. USER-BASED COLLABORATIVE FILTERING

A. User-based collaboratIve filtering

The user-based collaborative filtering algorithm measures and evaluates user preferences, including various related users, such as product information searches, reviews, product purchases, or collection of product data etc., to measure and evaluate these preferences [2]. Then calculate the attitudes and preferences of different users to the product or content, and finally use these data to calculate the relationship between different users, so that you can make mutual product recommendations for users with the same preferences. Figure 1 below shows the provision of recommended items for similar users. In the figure 1, the solid line with an arrow indicates a preference, while a dashed line with an arrow indicates a recommendation. The similarity between user 1 and user 3 is expressed. In Fig. 1, both user A and user C like item 2 and item 3, which means that their interests are similar, so item 1 and item 4 that user A likes are recommended to user C.

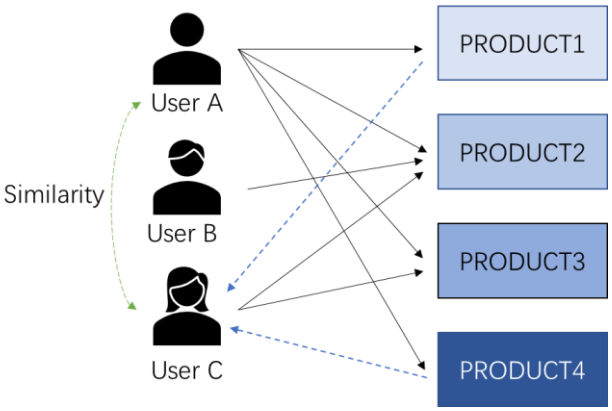

Fig. 1. User-based collaborative filtering recommendation algorithm

B. Look for similar users with preferences

For example, as shown in Table I, the ratings of five users on two products are shown to illustrate how to find similar users through users' attitudes and preferences for different products. In the example, 5 users rated two products separately. Then choose a correlation evaluation method to measure the correlation between users. Taking the Pearson correlation evaluation as an example, it will be 0.8-1.0 is a very strong correlation, 0.6-0.8 is a strong correlation, 0.4-0.6 is a moderate correlation, and 0.2- 0.4 weak correlation, 0-0.2 very weak correlation or no correlation.

Authorized licensed use limited to: ULAKBIM UASL - Osmangazi Universitesi. Downloaded on April 10,2025 at 19:27:19 UTC from IEEE Xplore. Restrictions apply.

TABLE I. PRODUCT RATING TABLE

	Commodity1	Commodity2
User A	3.5	6.8
User B	5.1	3.4
User C	4.2	2.8
User D	5.2	5.9
User E	6.5	4.2

C.Provide recommended items for similar users

When we need to recommend products to user C, we first check the previous similarity list and find that user C has a higher similarity with users D and E. In other words, these three users are a group and have the same preferences. Therefore, we can recommend products D and E to user C. But there is a problem here. We can't directly recommend the products of the previous product 1 to product 5. Because of these products, user C has browsed or purchased them. The recommendation cannot be repeated. Therefore, it is necessary to recommend products that user C has not browsed or purchased. User-based algorithms can only be used when there is a certain amount of data accumulation, that is, a cold start, because this algorithm needs to be calculated based on the user's historical data, so it cannot be applied to new users or newly created websites or software.

III. ITEM-BASED COLLABORATIVE FILTERING ALGORITHM AND MODEL-BASED COLLABORATIVE FILTERING ALGORITHM

In the collaborative filtering algorithm, the item-based and user-based algorithms are very similar. In these two algorithms, only the part of the user and the product is swapped. The project-based algorithm is to recommend similar products to users who may prefer this product, and the relationship between users is obtained by calculating users' different ratings for different items, and the scores are used to represent the user's attitude and attitude towards the product. Degree of preference. As shown in Figure 2, it shows the provision of recommended items for similar users. In the figure, the solid line with arrow indicates like, and the dotted line with arrow indicates recommendation. The similarity between user 1 and user 3 is expressed. Simply put, if user A purchases product 1 and product 4 at the same time, it means that product 1 and product 4 are similar. When the user C also purchases the product 1, it can be inferred that he also has a need to buy the product 4, and therefore recommends the product 4 to the user C.

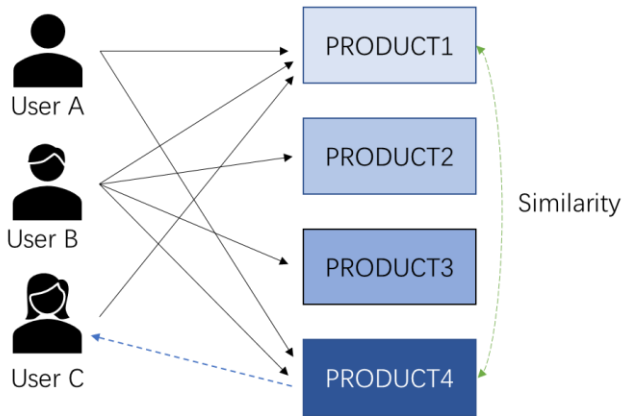

Fig. 2. Demonstration diagram of item-based algorithm

There is also a new model-based collaborative filtering algorithm. This model is used when the user or item data volume is particularly large when the collaborative filtering algorithm is used. Because the matrix in the calculation will be too large, model-based collaborative filtering appears. It mainly trains a model offline through technology such as machine learning, then optimizes the parameters of the model, and finally makes recommended predictions based on the user's preference information.

IV. COLLABORATIVE FILTERING SIMILARITY CALCULATION METHOD

The collaborative filtering recommendation algorithm mainly includes two collaborative filtering algorithms based on user and item-based. The following describes three evaluation methods for similarity calculation:

A. Euclidean Distance

Euclidean distance refers to the straight-line distance between two points in Euclidean space, and its associated norm is called the Euclidean norm, and the earlier literature is called the Pythagorean metric. [3] The formula is as follows:

$$d(x,y) = \sqrt{(\sum (x_1, y_1)^2)} \tag{1}$$

$$sim(x,y) = \frac{1}{1+d(x,y)} \tag{2}$$

B. Pearson correlation coefficient

Pearson correlation coefficient is used to measure the correlation between two variables X and Y [4], it is value of Pearson's correlation coefficient ranges from -1 to 1. When the value of the coefficient is 1, it means that all data points can be well or completely covered with a straight line, that is to say, X and Y can be well described by the straight-line equation to describe the relationship between them, and it is a positive correlation. When the value of the coefficient is -1, it means that although the straight line also contains all the data points, the time shows a negative correlation between them. If the value of the coefficient is 0, then there is no linear relationship between the two variables X and Y. The formula is as follows:

$$p(x,y) = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum xy_i)^2}} \tag{3}$$

C. Cosine similarity

Cosine similarity uses the cosine value of the angle between two vectors to express the similarity between them. When the two vectors point in the same direction, then according to the calculation of cosine similarity, the value is 1; when the two vector directions are at right angles, the value is zero; when the two vectors are pointing completely When the two directions are opposite, then the cosine similarity is -1. [5] It can be seen from this that when the cosine similarity calculation is used, the similarity value is only related to the direction of the two vectors, that is, the

Authorized licensed use limited to: ULAKBIM UASL - Osmangazi Universitesi. Downloaded on April 10,2025 at 19:27:19 UTC from IEEE Xplore. Restrictions apply.

size of the angle between the vectors. The formula is as follows:

$$T(\mathrm{x},\mathrm{y}) = \frac{x \cdot y}{||x||^2 \cdot ||y||^2} = \frac{\sum x_i y_i}{\sqrt{x_i^2}\sqrt{y_i^2}} \quad (4)$$

V. Comparative analysis of collaborative filtering algorithms

The article introduces several different algorithms for filtering collaboration. Although the calculation method of similarity is the same, the results based on different algorithms will be different. The user-based recommendation method mainly considers the calculation of the similarity between users, and the items recommended to the target user must be the items liked by the neighboring users. The item-based recommendation method uses similar items recommended to users with the highest ratings. Although the ideas of the two recommendation methods are different, they are both implemented based on the user item rating matrix. Therefore, in real applications, more need to consider application scenarios to implement algorithm selection. In some news websites, the number of users is far less than the number of news, and the item dimension is much higher than the user dimension. At the same time, users' interests and hobbies are relatively vague, and all kinds of news will be browsed. Therefore, choosing a user-based approach can greatly reduce space complexity. At the same time, the timeliness of news is relatively strong. If a project-based method is adopted, and news is the dimension of the project, it is unrealistic to continuously update the similarity of the project. At the same time, the goods are not very time-sensitive. Collaborative filtering applications usually involve very large data sets, while content-based methods only require the characteristics of the project information itself, rather than using interaction and feedback between users. Content-based recommendation will only recommend items similar to the items that the user has contacted before, and the diversity of recommendations in the recommendation evaluation index is very low. Since the new user does not have historical behavior data, the preferences of the new user cannot be obtained at this time, so there will be a user cold start problem.

VI. Conclusions

User-based methods and item-based methods have their own advantages and disadvantages. Content-based recommendation can solve the cold start problem of new users and data sparseness to a certain extent, but it cannot satisfy the diversity and novelty of recommendation. Recommendations based on collaborative filtering have data sparsity problems and low scalability. If the user tagging data is sparse, the item-based method is more accurate, and on the contrary, the user-based method is more accurate. In order to learn from each other and effectively integrate the two methods to improve the recommendation accuracy, a hybrid recommendation strategy was born, that is, a combination of multiple recommendation algorithm strategies, which can effectively avoid the shortcomings of a single algorithm, and ultimately produce a better recommendation effect than a single algorithm. At present, many effective hybrid models have been proposed in the academic community. The typical hybrid model idea is to use global parameters to linearly fuse the prediction results of the two methods. The recommendation system often needs to meet the recommendation requirements in various scenarios, and for different recommendation algorithms Fusion modeling can improve the robustness, diversity, and novelty of the recommendation system, and can improve the user experience while ensuring accuracy. Therefore, future research needs to study better hybrid strategies.

References

[1] Tao Sun, Kun Liu, Shaokai Zheng. "Design and Implementation of Public Opinion System in Colleges", Proceedings of the 2020 3rd International Conference on Big Data Technologies, 2020

[2] Li Hui Nian, Jing Wei, Can Bin Yin. "The promotion role of mobile online education platform in students' self-learning", International Journal of Continuing Engineering Education and Life-Long Learning, 2019

[3] Tingting Yu, Yimin Hu, Guanping Feng, Ke Hu. "A Graphene-Based Flexible Device as a Specific Far-Infrared Emitter for Noninvasive Tumor Therapy", Advanced Therapeutics, 2020

[4] Wenwei Dong, Yanlu Xie, Binghuai Lin. "Unsupervised Pronunciation Fluency Scoring by infoGan", 2019

[5] Minxuan Li. "An improved FCM clustering algorithm based on cosine similarity", Proceedings of the 2019 International Conference on Data Mining and Machine Learning - ICDMML 2019, 2019

Authorized licensed use limited to: ULAKBIM UASL - Osmangazi Universitesi. Downloaded on April 10,2025 at 19:27:19 UTC from IEEE Xplore. Restrictions apply.