Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/4/2023371

Operation and algorithm optimization of short video recommendation algorithm

Yilin Wu

Wuxi No.1 High School, Wuxi, 214044, China

Willing1010510@163.com

Abstract. Now that technology is developing rapidly, the field of computing has been showing exponential progress, and scientific and technological civilization of mankind is making continuous progress. Games, live broadcasts, and short videos are all new products under big data. Especially for short videos, the huge database records the preferences of billions of users. With the combination of collaborative filtering algorithms and content-based recommendation algorithms, computers can always accurately recommend suitable videos to various users. In this paper, improvement of this method is aimed at the item-based algorithm in the collaborative filtering algorithm. For the item-attribute matrix, first, use the Jaccard distance to calculate the similarity, and then use this similarity value instead of the Euler distance formula to bring it into the k-means clustering, and use iteration to obtain countless different clusters. Finally, set a threshold x, which is the distance between each cluster center. Whenever there is a new matrix to be classified, the similarity y corresponding to this matrix is calculated first. If y<x, the matrix is classified into the corresponding cluster. This approach can improve the diversity of recommended videos and tap the potential interests of users. Such improvements to the matrix can improve the accuracy of the algorithm and user stickiness.

Keywords: Collaborative Filtering, Recommendation System, K-means, Jaccard Distance, Clustering, Matrix.

1. Introduction

Since the emergence of the recommendation system, it has gone through several decades of development history. In the early recommendation system, the products are usually simply sorted according to their popularity of the products, and several items at the top of the ranking are directly recommended to the user. The method also has a good effect to a certain extent, and it is still used in the current recommendation system. However, this method has certain limitations.

Now big data runs through human life, especially for young people, everyone's life is inseparable from all kinds of electronic products. Among them, games and short videos are the most attractive. Young people's entertainment time is all devoted to these. There are tons of videos available on Youtube, Twitter, Tiktok. People are passively recommended video after video that interests them, which results in no one putting down their phone easily. Through the calculation of big data, computers can even provide people with potentially favorite videos, which is addicting. As a non-human actor, the recommendation algorithm forms a short video content distribution network together with short video platforms, short video content producers and operators, and users. Different

DOI: 10.54254/2755-2721/4/2023371

short video companies will use different recommendation algorithms. This is because each company wants its algorithms to be more accurate at capturing user preferences for recommendations. The more accurate a company's recommendation algorithm is, the more competitive the company is. A variety of recommendation algorithms are essentially a combination of collaborative filtering algorithms and content-based recommendation algorithms. For collaborative filtering systems, it is essentially a giant user-item matrix. Based on the excessive sparsity of this matrix, different algorithms of major companies are dedicated to completing this matrix with extremely low accuracy. In fact, in the face of massive data and users, all matrix completion methods are not particularly effective. Therefore, the method proposed here is mainly based on the improvement of another type of collaborative filtering algorithm—item-based collaborative filtering algorithm. This improvement can not only greatly improve the problem of recommendation accuracy, but also explore the potential interests of users and recommend videos that users may be interested into them. This adds some randomness and probability to the algorithm.

2. Basic idea and classification of collaborative filtering

2.1. What is collaborative filtering

Collaborative filtering first appeared in the news recommendation system of Group Lens. At present, commonly used personalized recommendation algorithms include a content-based recommendation, collaborative filtering-based recommendation and hybrid recommendation. The collaborative filtering algorithm is simple, intuitive and easy to implement, so it is the most widely used. A collaborative filtering algorithm refers to a recommendation algorithm that uses the behavior information of other users to infer a user's interests and preferences [1]. In the early days, it was mainly divided into user-based collaborative filtering algorithms and item-based collaborative filtering algorithms [2]. The calculation process is to use a certain behavior of the user to the item to construct a user-item behavior matrix, and obtain the user vector or item vector from the matrix to calculate the similarity between the vectors to obtain the user-user similarity matrix or item-item similarity. matrix, and then interact with the behavior matrix to calculate the user's recommendation score for each item. Due to the need to calculate the similarity between each user or each item, the computing performance becomes a big bottleneck, but the actual interaction between users and items is only a small part of them, so a considerable amount of calculation is unnecessary In addition, the calculation of collaborative filtering requires interaction between users and items, which is unfair to new users and new items, making it difficult for the recommendation system to generate accurate recommendations for them.

2.2. User-Based Collaborative Filtering (UBCF) and Item-Based Collaborative Filtering (IBCF)

The algorithm is based on the collective intelligence of humans as social animals. The most widely used collaborative filtering algorithm is user-based collaborative filtering. As its name suggests, the essence of this algorithm is to recommend similar videos by finding similar users. As shown below (Figure 1), suppose there are now U1, U2, U3, U4, four users. U1, U2 and U4 all like Item1, Item2, Item4 (Video1, Video2, Video4). Therefore, the computer will think that U1, U2 and U4 belong to user type A with similar interests. Since U2 and U4 like Item4 together, users of type A are likely to like Item4, and the computer will also recommend Item4 to U1, so as to achieve the purpose of recommending videos based on user similarity. Such a recommendation still has drawbacks, that is, the measurement of similarity is too simplistic. And such an inference algorithm can only work when the user provides a large amount of information. The huge data sparsity problem and the cold start problem are all stumbling blocks for this algorithm.

DOI: 10.54254/2755-2721/4/2023371

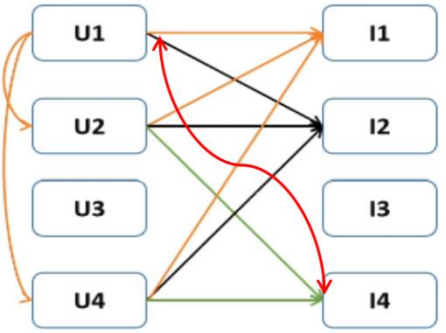

Figure 1. User-based CF.

Another type of collaborative filtering algorithm is an item-attribute matrix based filtering algorithm. As shown in the figure below (Figure 2), there are still four users U1, U2, U3, U4. This time U2 and U4 both like Video2 and Video3. At this time, the computer will think that Item2 and Item3 are similar, and U1 also likes this time Video2, then Big Data will recommend Video3 to U1. Similarly, this algorithm also has flaws, and the video recommendations that users get are not personalized. There is also no way for the system to mine videos that users are potentially interested in through this algorithm.

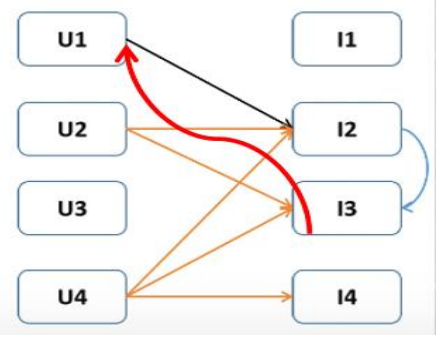

Figure 2. Item-Based CF.

3. The improvement of collaborative filtering

3.1. Disadvantages of sparse matrices

Although collaborative filtering algorithm's scope of application is very wide in the computer However, due to the ever-increasing number of users, video screens, and a large base of existing users, short video screens, and the users have only scored a few products, the score matrix contains a huge number of missing items, that is, the problem of data sparsity. However, the inaccuracy of similarity measurement caused by data sparsity is becoming more and more prominent, which directly leads to the rapid decline of recommendation quality.

As shown in the figure below, with hundreds of millions of users and hundreds of millions of videos, we will get a very huge matrix. If a user U_1 interacts with item 1, then R_{11} is 1. If U1 does not interact with item I_2, then $R_{12}$is 0. From this derivation, it can be obtained that most of the positions are recorded as 0, and the dimension is extremely Large rectangles are extremely sparse. This is a difficult problem that user-based collaborative filtering algorithms must face.

Therefore, matrix decomposition methods such as SVD and NMF have been applied in the recommendation system, and have achieved corresponding results. Later, some improved matrix decomposition models have emerged, such as PMF, SVD. The model-based collaborative filtering algorithm has gradually occupied the dominant position in the application market. After that, more and

DOI: 10.54254/2755-2721/4/2023371

more models are being used in recommendation systems, including the most popular deep learning recommendation model, which is also deeply influenced by the idea of collaborative filtering.

However, since collaborative filtering algorithms are not limited to user-based recommendation algorithms, this paper hopes to make breakthroughs from item-based recommendation algorithms to solve the difficulties of lack of diversity and personalization. At the same time, it further improves the accuracy of the collaborative filtering algorithm itself.

Table 1. Short video preference chart.

User/ Item	I_1	I_2	...	I_j	...	I_n
U_1	R_{11}	R_{12}	...	R_{1j}	...	R_{1n}
U_2	R_{21}	R_{22}	...	R_{2j}	...	R_{2n}
...	...	...	...	...	...	...
U_i	R_{i1}	R_{i2}	...	R_{ij}	...	R_{in}
...	...	...	...	...	...	...
U_m	R_{m1}	R_{m2}	...	R_{mj}	...	R_{mn}

3.2. Measurement of similarity

Since the improvement of this algorithm needs to accurately calculate the similarity between two matrices, Two similarity algorithms are compared here. As shown in the figure below, this table is the data obtained from the experiment I found seven different types of videos and four friends. For these seven different videos, I randomly asked my friends to watch a random number of videos and asked them to rate what they saw. The score is 1-5, with 5 being the favorite and 1 being the dislike. Friends A, B, C, and D all gave their corresponding scores respectively. Four sets of corresponding user vectors are obtained.

Table 2. User-item Matrix.

	I_1	I_2	I_3	I_4	I_5	I_6	I_7
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

The first similarity calculation method is the Jaccard similarity calculation. Use the formula:

$$sim(A,B)= \frac{|r_A \cap r_B|}{|r_A \cup r_B|}$$

And so on to get the similarity between friends B, C, D and friend A.

figure out $sim(A,B)=\frac{1}{5}$, $sim(A,C)=\frac{2}{4}$, $sim(A,D)=\frac{0}{5}$,

It is calculated that friend C is the most similar to friend A, and friend D has the lowest similarity. But looking at the graph data again, we can find that both A and C have evaluated I4 and I5, but their

DOI: 10.54254/2755-2721/4/2023371

evaluations are quite different. For I4, friend A gave a high score of 5 points, but C only gave 2 points. I5, friend A only gave 1 point, but friend C gave a high score of 4 points. It can be seen that the similarity between A and C is not high, and their choice of video is very different. From this, it can be concluded that the defect of this algorithm is that it does not take into account the specific numerical situation of the video score. This similarity calculation only uses the number of videos watched by the user, and the score does not play a decisive role.

The second similarity algorithm is the cosine similarity algorithm, that calculates the angle between two corresponding vectors to replace the so-called "similarity". $sim(A,B)= \cos(r_A,r_B)=\cos \frac{\vec{x}\cdot\vec{y}}{|\vec{x}|\cdot|\vec{y}|}$, obtained by calculation sim(A,B)=0.38, sim(A,C)=0.32, It can be seen from the calculated values that the similarity difference between friends A and B, and friends A and C is not large, so it cannot reflect the phase velocity difference and similarity distance well. This is because all vacancies in the matrix are automatically assigned a value of 0 in this algorithm, but these vacancies actually represent that this friend has not seen this video. However, 0 is a lower value than 1, and the algorithm directly defaults that the videos the user has not watched are their least favorite. And this is illogical. In the face of massive data that needs to be calculated, this algorithm will bring low-level accurate values[4].

3.3. K-means clustering

Because the user-based collaborative filtering algorithm relies heavily on the calculation of similarity and the decomposition of huge dense and dense matrices. Therefore, in terms of improvement, this article chooses to improve the item-based collaborative filtering algorithm. As shown below, this is an item-attribute matrix. Compared to the user-item matrix, the dimension of this matrix is greatly reduced and becomes more compact. Therefore, this article proposes to perform a clustering of the similarity matrix before recommending videos. As shown below，I_1represents Item 1，$a_1$represents attribute1, If item 1 has the attribute of attribute 1, it is recorded as 1, and the actual value on the matrix is 0, which means that item 1 does not have the characteristics of attribute 1.

Table 3. Item-attribute Matrix.

	a_1	...	a_j	...	a_s
I_1	0	...	0	...	0
...	...	...	...	...	...
I_j	1	...	1	...	1
...	...	...	...	...	...
I_s	1	...	0	...	0

Based on the item attribute feature matrix, K-means algorithm is helpful and necessary for clustering items. Traditional K-means algorithm uses the normal Euclidean distance to measure the attribute distance of two items, and generates new cluster centers through continuing iteration till the clusters are stable. Because the item attribute matrix is Boolean data, the Euclidean distance cannot well represent the difference of attributes between items, and it is quite hard to obtain the new cluster centers by averaging during iteration [3]. =Therefore, based on the K-means algorithm, this paper performs the following steps. Therefore, instead of using Euclidean distance, the similarity measured by jaccard distance is used instead of Euclidean distance. The calculation formula is:

$$d_j(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} = \frac{M_{10}+M_{01}}{M_{10}+M_{01}+M_{11}}$$

DOI: 10.54254/2755-2721/4/2023371

$M_{11}$is the number of attributes whose attribute value is 1 for both items; M_{01} is the number of attributes whose attribute value is 0 for one item and 1 for another item; $M_{10}$is the number of attributes whose attribute value is 1 for one item and 0 for the other item. In the iterative process of clustering, the new cluster center does not use the numerical average method, but takes the item with the smallest attribute distance from the same item as the cluster center. Suppose now I have n clusters, which are C_1,C_2,.....C_n, Calculate the attribute distance between each item in the class and the items of the same cluster, then we can get$d_{i,1}$, $d_{i,2}$,.....$d_{i,n}$. As a result, sum of the attribute distance is d_i= $d_{i,1}$+$d_{i,2}$+.....+$d_{i,n}$. Take the item with the smallest attribute distance and D_i as a new center, complete the update of all cluster centers, and finally iterate until the cluster is stable. The algorithm is more complex, but it can be done by the fast-calculating computers, so it does not matter. Then we can get n clustering center (c_1,c_2,.....c_n) and their accordingly clusters (C_1,C_2,.....C_n).

3.4. The given threshold

If only searching in the same category, you can only recommend items with similar attributes to users, which lacks novelty and makes it difficult to tap users' potential interests. Therefore, on the basis of clustering, this paper sets the item attribute threshold, filters out the categories whose attribute distance is within the threshold range, and searches for neighbors in the filtered categories. [5] Based on this, I will first set a threshold δ,which should be determined according to the actual situation. Next,I will figure out the attribute distance between our target item and all cluster centers,let attribute distance of the i-th item as $d^a(i)$=($d_1, d_2, \ldots\ldots, d_n$). Finally, to all the attribute distance that away from target item i ($d_1, d_2, \ldots\ldots, d_n$), if this $d_j<\delta$, $j\in 1,2,3,\ldots n$, Then cluster j is classified as the neighbor search range S(i) of target item i. Based on this, the search range will be largely narrowed, and the calculation time is correspondingly reduced, so the efficiency of the algorithm is improved to a certain extent.

4. Conclusion

Here introduces the most basic collaborative filtering algorithm and similarity calculation method, and also puts forward their defects. Aiming at the defects of low efficiency and low accuracy of traditional collaborative filtering algorithms, this paper proposes a collaborative filtering algorithm based on item attribute clustering and similarity optimization [6]. First, the similarity calculation method is improved by combining the user's subjective rating and the item's objective attributes, so as to avoid relying too much on the user's subjective rating; then, the search scope is narrowed and the computational complexity is reduced.

References

[1] LIU Duan-yang. Design and implementation of video recommendation system based on deep viewing interest network. China Academic Journal Electric Publishing House, 2021.6.1, 18,19

[2] LIU Xian-feng, LIU Tong-cun. Research on project grading prediction and recommendation algorithm based on attribute clustering, 2021,9,10,11

[3] SU Kai, ZHANG Xuan, FU jin. Collaborative filtering algorithm based on attribute clustering and similarity optimization. China Academic Journal Electric Publishing House, 2021,11,30, 3,4

[4] SUN Hong-mei. Research on Optimization of Collaborative Filtering. Algorithm.China Academic Journal Electric Publishing House,2021.5.1,1

[5] WEN Feng-ming, XIE Fang-xue, Operational logic and ethical concerns of short video recommendation algorithms,China Academic Journal Electric Publishing House,2021.8.6,2

[6] GU Ming-xing, Huang Wei-jian, Huang yuan.Collaborative filtering recommendation combining attribute clustering and improving user similarity,2020, 187,188,189