Academic Journal of Computing & Information Science
ISSN 2616-5775 Vol. 4, Issue 1: 94-104, DOI: 10.25236/AJCIS.2021.040115

A Recommendation Algorithm Based on Efficient Dense Subsequence and Data Generation

Li Zhitao[a], Zhou Xiaokang[b]

Department of Software Engineering, Nanjing University of Finance and Economics, Nanjing, Jiangsu, China
[a]E-mail: 2997017886@qq.com, [b]E-mail: 1990858822@qq.com

Abstract: *[purposes] The recommendation effect of the user-based collaborative filtering algorithm is improved by extracting the effective dense sub-sequence between the user and the project and generating the data information of the project that the user participates in with high probability, The problem of data sparsity is alleviated to some extent, and the validity of binary data method in expressing user interest is verified. [methods]There are two steps to extract the effective dense subsequences from the user rate sequence, Firstly, the effective dense subsequence is extracted based on the time range; secondly, the subsequence is extracted twice by judging the items in the subsequence with high probability that do not conform to the user's interest. Generating data for projects with high probability of user participation requires the use of project attributes of user participation projects. The validity of the binary data method is mainly verified by the comparison experiment with the scoring data and the final result of the recommendation algorithm. [results] The experimental results show that RMSE (root mean square error) is reduced by more than 0.04 on the ml-latest-small data set compared with the classical traditional collaborative filtering algorithm. When F value (the harmonic mean of accuracy and recall rate) is taken as the evaluation standard, the algorithm proposed in this paper is more accurate than other classical algorithms, and the recommendation accuracy has been significantly improved. [conclusions] The method of extracting the effective dense subsequence between the user and the project is better than the traditional collaborative filtering method in terms of recommendation effect, and using binary data in the algorithm to replace the original rate data can bring better recommendation effect.*

Keywords: *Dense Subsequence, Binary Data, Generate The Data, Mitigation of Sparsity, Recommendation System Algorithm.*

1. Introduction

The application of recommendation system in our daily life has become so common that some scholars assert that "we are leaving the information age and entering the recommendation era"[1]. At present, the recommendation system has been widely used in deep learning[2]/e-commerce[3]/digital library[4]and other systems. More and more competition between e-commerce sites and social media has gradually turned into competition for personalized recommendation services. The purpose of the recommendation system is to judge the possible points of interest of users in the future based on the existing user interest history, so as to recommend the potentially interesting content that users have not yet paid attention to. Therefore, it is a major challenge for the recommendation system to recommend items more in line with users' interests according to users' browsing information or purchasing items in real time.

After years of research and development, the algorithm improvement of the recommendation system has benefited from the contributions of many different disciplines, among which the more common and important disciplines include computer science, statistics, information science, etc. However, at the same time, we also noticed a very noteworthy trend of research, that is, the algorithm to improve the recommendation system cannot be completely based on the improvement of the algorithm itself, as the key content of all the current recommendation system algorithms -- the data itself, has not received too much attention. At present, most scholars believe that rating behavior is a very effective way to judge users' interest, and the existing research methods are mostly based on this hypothesis. Clearly, this scoring behavior and its information is a common type of explicit information that can be provided by various common recommendation system test datasets.

ISSN 2616-5775 Vol. 4, Issue 1: 94-104, DOI: 10.25236/AJCIS.2021.040115

However, existing experiments have shown that recommendation algorithm performance is affected by selection data, and even changing the order of project elements presented to users can have a significant impact on the recommendation effect. For example, displaying items by popularity can improve RMSE by 5.5%[5]. For another example, the variance of the rate can be divided into the variance of the rate calculated according to the user and the variance of the rate calculated according to the project. Generally speaking, when the variance of project rate increases, the accuracy of recommendation system will decrease[6]. At the same time, this kind of influence is not limited by the recommendation algorithm and exists widely. Therefore, the accuracy of the recommendation system can be effectively improved by using the rate data of small item variance[7].

Therefore, it provides a beneficial research direction for the algorithm improvement of the existing recommendation system, that is, how to effectively select data to express more accurate user interest pattern, and further proposes the idea of improving the algorithm effect of the existing recommendation system. This paper combined with the standard collaborative filtering method to make a preliminary study attempt. Collaborative filtering is a very effective and widely used personalized recommendation technology[8], It is based on the simple assumption that a user's past interests are a proxy for his future interests. Therefore, by analyzing the existing user interest information, which is mainly reflected by the user's rating of the project, the rate of unknown project in the future can be predicted. This assumption is reasonable to some extent. For example, some scholars analyzed the user data on the recruitment information website and found that, for each user, 2/7 projects in the past 14 weeks were clicked again by the user in the 15th week on average[9].

Specifically, collaborative filtering recommendation method is to provide the current users with appropriate suggestions or projects by obtaining other users similar to the current users. The advantage is that you don't need to know the specific content of the project, but you can also recommend new projects that users may be interested in. However, the traditional collaborative filtering algorithm also has many shortcomings: first, the traditional collaborative filtering algorithm cannot handle data that is too sparse, so it is easy to generate inaccurate similarity calculation; Secondly, whether the rate used in the traditional collaborative filtering algorithm can effectively express user needs has not been accurately verified.

For this reason, this paper focuses on the effectiveness calculation of the collaborative filtering recommendation algorithm based on effective dense sub-sequences, that is, by reasonably selecting effective dense subsequences in the existing user scoring records, the density of effective data is increased and the adverse interference of noise data is reduced. However, at the same time, it will bring more serious data sparsity problem to the source data. In order to address this situation, this paper generates projects with high probability for users by analyzing the attribute characteristics of projects and combining with the effective time interval of users, The effective time interval refers to the starting time for users to participate in project evaluation to avoid excessive migration of users' interests over time. At the same time, this article also uses the user interest changes to in-depth study and analysis of the user ratings information with the user whether the effect of the evaluation in terms of validating the user demand, according to an evaluation of whether the user of binary data to replace the traditional specific rating numerical data expression method, experiment shows that the application of user evaluation binary data will be more excellent results.

2. Related Work

The traditional collaborative filtering algorithm mainly includes three important steps, namely obtaining data, searching for nearest neighbor elements, and predicting and recommending. Among them, finding the nearest neighbor element often has the most fundamental influence on the algorithm of collaborative filtering recommendation system, which can directly influence the final recommendation effect.

The traditional collaborative filtering algorithm is difficult to solve the problem of data sparsity, literature[10-12] proposes different solutions to this problem. In order to solve the problem caused by data sparsity of collaborative filtering recommendation system, literature[10] uses sigmoid function to realize smooth transition of user attributes and user rate information in user similarity calculation under data sparsity. Literature[11] also use sigmoid function to process the data on the user's rate is data sparseness, literature[12] proposed a weighted fusion sparse degree of collaborative filtering algorithm to solve the data sparseness, redefined the sparse matrix in the algorithm to calculate method, and then the fusion weighted matrix sparse degree of similarity to the user, and to improve the collaborative

ISSN 2616-5775 Vol. 4, Issue 1: 94-104, DOI: 10.25236/AJCIS.2021.040115

filtering algorithm. This method can indeed solve the sparsity problem, but it aims to improve the similarity. This paper deals with the sparsity problem mainly by generating project data with high probability for each user, which to a large extent makes up for the shortcoming that the source data is too sparse. In order to effectively mine the changing trend of users' interests, literature[13-14] proposed a collaborative filtering recommendation algorithm based on the changes of users' interests. Literature[13] observed users' interests through the forgetting curve to adapt to the changes of users' interests. Literature[14] further weighted the similarity results by using the user's interest proximity degree, and the obtained similarity results incorporated the user's interest and preference information. On this basis, collaborative filtering algorithm was adopted to carry out personalized recommendation.

This paper extracts effective data on the source data for each user to provide the most valuable dense dataset that will be improved to some extent after extraction of source data sparseness degree, this paper generated high probability for each user to participate in the project to solve the problem of sparse data, is more than makes up for the extraction of dense subsequence, also eased the source data to data sparseness of the data itself. The study of user interest involves whether the user evaluates or not as the simplest form of implicit information. We can set a binary value of whether the user has rated an item, and then represent pseudoimplicit rating. The relevant reason is that the user's evaluation of the project is not a random behavior, and the evaluation behavior itself reflects a kind of user's preference information for the project[15]. Even though this information is not sufficient, previous studies have shown that integrating implicit information into existing explicit information can increase the prediction accuracy of the recommendation system compared with the method that simply uses explicit user information[16]. Based on the in-depth study of users' interest changes in many literatures[17-21], this paper also makes a detailed analysis of users' interest changes.

3. A Recommendation Algorithm Based on Efficient Dense Subsequences and User Potential Interest Analysis

3.1 Extraction of Effective Dense Subsequences in User Scoring Sequences

By observing the traditional collaborative filtering algorithm, we can clearly find that in the constructed scoring matrix, most of the users are not involved in the project, leading to unusually sparse data. This is a very important reason why this paper proposes to effectively extract the dense subsequence of the project information. In order to alleviate the hard to avoid problems, the paper proposed a method of extracting effective dense subsequence to improve this problem, to extract effective dense subsequence is divided into two steps,

Step 1: the first is through time factor carries on the preliminary subsequence extraction, after extracting the data more dense than the source data, and data to be more effective.

Step 2: Filter the subsequences obtained in the first step to filter out the project data that does not meet the user's interest but the user actually participates in. At the same time, a better prediction model of the recommendation system is proposed.

3.1.1 First Extraction

Assuming that the earliest and the latest projects evaluated by user u were in 2012 and 2019, and the earliest and the latest projects evaluated by user v were in 2010 and 2015, the earliest and the latest evaluation time of user u and user v were selected as 2010 and 2019, respectively. All the scoring items between [2010, 2019] were selected to form a new molecular sequence, based on which the similarity between user u and user v was calculated. It should be mentioned that different users get different time intervals.

3.1.2 Second Extraction

The purpose of the second extract is more accurate to find out the user's interest, because users to participate in a project does not have to because I like the whole content of the project, part of the project is likely to be like, and when the user to the evaluation of this project is very low, that don't like this project, this paper will filter out this part of the project. It can be analyzed that this part of the project is mainly concentrated in the low evaluation set. Therefore, filtering out the data in line with the characteristics of this part of the project in the low evaluation project is the core of the second extraction. The specific methods are as follows:

Combined for the first time to extract the subsequence, statistical evaluation items of each user's

ISSN 2616-5775 Vol. 4, Issue 1: 94-104, DOI: 10.25236/AJCIS.2021.040115

low attribute characteristic distribution of selected attributes features the most sparse distribution properties, the characteristics of the filter out low project attributes contained in the low distribution of the characteristics of the project, the resulting sequence contains two parts: user real like low project set and high project set. The reason for the low rate can be completely attributed to the quality of the project itself.

3.2 The Data Generated

Aiming at the problem of increasing the sparsity of source data caused by III) A) data extraction, this paper conducts data generation for projects that users participate in with high probability, which not only makes up for the defects of data extraction, but also alleviates the situation that the data sparsity of source data is too high. Data generation mainly relies on low-rated projects. Projects with high frequency of user participation in low-rated projects may be movies that users like, because even with low rates, users still participate in such projects all the time, indicating that users are really interested in such projects. The purpose of data generation is to find this part of the project for each user and generate it, specifically as follows:

According to the collection of each user and project, find out the project attribute type with the highest cumulative occurrence times in the low-evaluation project. If there is a situation with the highest joint occurrence, select all of them. There will be many such situations in the source data. Another case, the user's behavior may be not active, involved in the project is relatively small, lead to project the characteristics of the property type statistics will be very few, therefore defines a constraint, when there is the highest cumulative project attribute type number more than half the number of users to participate in the project, which means the user participate in low evaluation project, more than half are this type of project, select the user to participate in the project starting time attribute types of projects, all these projects using the average evaluation standard of the user data is generated for this kind of project evaluation. When the generated data overlaps with the data in the source data during data generation, the real data in the source data is retained.

3.3 Improved Similarity Calculation Method and Its Recommendation Algorithm

In this paper, not only extract the effective dense user rating molecular sequence, but also compare the expression of binary data and rating data. Among them, the similarity calculation for traditional collaborative filtering methods can not deal with binary data sequence similarity comparison, such as using cosine similarity to calculate, will cause the denominator of 0 nonsense, and use the adjusted cosine similarity and Pearson[24] relevant similarity calculation inevitably need to calculate the average rate, for binary data, the average doesn't make any sense. Therefore, referring to[25], function(1) is used to calculate similarity:

$$\mathrm{sim}(I,J)=\frac{(I\times J)}{(\||I|\times|I\|+\||J|\times|J\|-I\times J)} \tag{1}$$

$$\mathrm{sim}(\mathrm{I},J)=\frac{I\cdot J}{\|I\|\cdot\|J\|} \tag{2}$$

In functions(1) and (2), I and J are the scoring vectors of two different users, and the calculation results of function (1) are constrained between the interval [0,1], which is more convenient for calculation in the later stage of the algorithm than the result interval of Pearson's correlation similarity [-1,1]. The function (2) is a vector representation of the cosine similarity.

After the similarity calculation, we can take the result with a high similarity as the nearest neighbor set. The calculation method of the predicted value is shown in function(3):

$$P_{a,\mathrm{i}}=\overline{r_a}+\frac{\sum_{b\in N}\mathrm{sim}(a,b)(r_{b,\mathrm{i}}-\overline{r_b})}{\sum_{b\in N}sim(a,b)}（\mathrm{i}\in L） \tag{3}$$

Where, L is the final data set obtained after III) A) and III) B) of the training set data of the source data, $\overline{\mathrm{r}_\mathrm{a}}$ is the average of user a ratings of all items, $\overline{\mathrm{r}_\mathrm{b}}$ is the average of user b ratings of all items,

ISSN 2616-5775 Vol. 4, Issue 1: 94-104, DOI: 10.25236/AJCIS.2021.040115

$r_{b,k}$ is user b rating of project k, $\mathrm{sim}(a,b)$ is the similarity between user a and user b, N is a nearest neighbor set.

The procedure of the specific algorithm is described as follows:

Algorithm: Collaborative filtering algorithm based on efficient dense subsequences and user potential interest analysis.

Input: The data set is divided into a pair of training set and test set, and the nearest neighbor is N.

Output: The predicted value of user a project i rate in the test set.

Detailed steps of algorithm:

Step 1: Extract effective dense subssequences for each user in the training set using method III) A), then generate data for each user using method III) B), and then generate corresponding project scoring matrix for each user, For binary fillings: fill with "1" if there is a rate, and fill with "0" if there is no rate. For those who use the rate value method to fill, just use the rate to fill.

Step 2: The binary value method user function(1) to calculate the user similarity between user a and user b, and the scoring value method uses function(2) to calculate the user similarity between user a and user b.

Step 3: According to the similarity calculated in step 2 users to find the nearest neighbors, assume that a nearest neighbors for user a, the user first a similarity with other users results from high to low order, then according to the need to choose appropriate similarity result row in front of the user as a collection of nearest neighbors, and function(3) is used to calculate the user a focus on training project grading forecast.

4. Validity Study of Binary Data on Whether the User Rate

It has been explained above that the binary data information whether the user has evaluated can be understood as a potential user interest, that is, the project selected by the user solely based on his interest can represent the user's potential interest before the project is graded by the user. For example, in real life, people usually go to see a movie not because they hear that it is very good, nor because the movie has a high rate. More often than not, they choose to watch the movie because it is their favorite type. Because of this, when two people choose the same movie because of their respective potential interests, the results of other things can be predicted more accurately through mutual analysis.

Take the example of film data set, the data in the traditional collaborative filtering algorithm of rating matrix is the user value of film rate, the rate values can be said to be in the user after watching this movie, given to the evaluation of the film, if the rate of 4 or 5 points, can be thought of as the user like the film, also can be thought of as the user is out of love for this kind of movie, or the users are just like the star of the film, the reason is vary from person to person, want to put these many reasons together, the workload is very huge and difficult to achieve. Therefore, based on whether the user evaluation of binary data expression, users can provide a research only innate interest in simple way, don't need to pay attention to the user how many points called a film, also do not need to pay attention to what the user is based on the reasons for film rates, just need to focus on the user ever watched the movie, if you've seen, the token user and the relationship between the film as "1", or "0".

However, the idea of this new data expression needs experimental verification, so the following experiments are designed:

The experimental method: To compare the effectiveness of binary data and rating data in expressing users' interest by comparing the similarity of users' past and future rating items.

The experimental steps:

Step 1: Divide each user's evaluation items into training set and test set according to the time sequence of scoring. The data in the training set is the items that the user evaluated in the past, and the data in the test set is the items that the user will evaluate in the future.

Step 2: use the number of evaluation item types of each user in the training set to form a vector. For example, for user 1, the vector formed in the training set is [5,20,6,4,8...]. , the total number of 19 item types, and the sequence formation vector in the test set was found in the same way [8,12,7,2,16...].

Step 3: Calculate the similarity of the sequence obtained by each user. The method of similarity calculation uses the function(1) in III C, and finally take the average value of the similarity of all users.

5. Instructions of Experiment

5.1 Dataset and Evaluation Methods

In this paper, the ml-latest-small dataset is selected, and the data structure is shown in Table 1:

Table 1: Structure table of dataset.

Dataset	Size	User	User range	Item	Item range	The amount of data
ml-latest-small	2771K	610	[1,610]	9724	[1,193609]	100000

The evaluation criteria of recommendation effect are analyzed from RMSE (root mean square error) and K value (harmonic average of accuracy and recall rate). The calculation functions(4), functions(5) are given below:

$$\mathrm{RMSE}(X,h)=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(x_i)-y_i)^2} \quad (4)$$

In function (4), X can be understood as a set of users, and the function $\mathrm{h}(x)$ is a recommendation system model, y_i is the real evaluation data of each user for the project.

$$F = accuracy * recall * 2 / (accuracy + recall) \quad (5)$$

The accuracy and recall are calculated as follows:

accuracy = The number of correct messages extracted/Number of extracted messages

recall = The number of correct messages extracted/Number of messages in the sample

The calculation formulas for accuracy and recall of different dataset are also redefined with different data set structures. The data structures adopted in this paper are shown in Table 1. Therefore, they are redefined as follows:

Number of correct messages: the total number of data that the absolute value difference between the predicted data and the real data is less than 0.5.

Number of extracted messages: the total number of data whose predicted data value is higher than 3.

Number of messages in the sample: The total number of data whose real data value is higher than 3.

5.2 Validity of Dataset Partitioning

Since the experiments in this paper all involve the partition of dataset, if the partition of dataset is not reasonable, it will not only cause the dispute of algorithm results, but also reduce the robustness of the algorithm to some extent. So, before experiment, verify the validity of the data set classification is particularly important, this article divide the data set is based on user participation in time sequence of the project, participated in project evaluation in the earliest users as the beginning of the training set, the following is given according to the different proportion of division of ml - the latest - the condition of the small (including similarity index using function (2)):

Table 2: Rate distribution (60%training,40%test).

60%training, 40%test	Rate range	Training data volume	Test data volume	Similarity
1	(0,1]	3425	2685	0.99840
2	(1,2]	6517	4853	
3	(2,3]	16104	11041	
4	(3,4]	20808	13366	
5	(4,5]	12765	8436	

ISSN 2616-5775 Vol. 4, Issue 1: 94-104, DOI: 10.25236/AJCIS.2021.040115

Table 3: Rate distribution (50%training, 50%test).

50%training, 50%test	Rate range	Training data volume	Test data volume	Similarity
1	(0,1]	2846	3264	0.99856
2	(1,2]	5387	5983	
3	(2,3]	13436	13709	
4	(3,4]	17348	16826	
5	(4,5]	10743	10458	

Table 4: Rate distribution (40%training, 60%test).

40%training, 60%test	Rate range	Training data volume	Test data volume	Similarity
1	(0,1]	2245	3865	0.99861
2	(1,2]	4217	7153	
3	(2,3]	10716	16429	
4	(3,4]	13861	20313	
5	(4,5]	8594	12607	

Based on the above analysis of the scoring data after dividing the data set, after dividing the data set by different proportions, the vector similarity of the number of scores in the scoring interval before and after is up to 99%, which is more in line with the interests of users. Therefore, it is reasonable and effective to divide the data set by the time users participate in the project in this paper.

5.3 Results Analysis

5.3.1 Comparison Experiment of Validity between Binary Data and Scoring Data

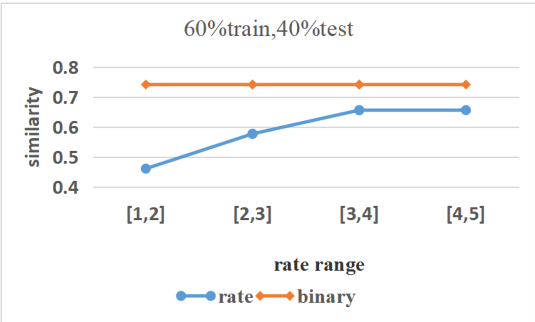

Figure 1: Divide 60% of the source data into training sets and 40% into test sets, using binary data and scoring data based on the experimental results in IV.

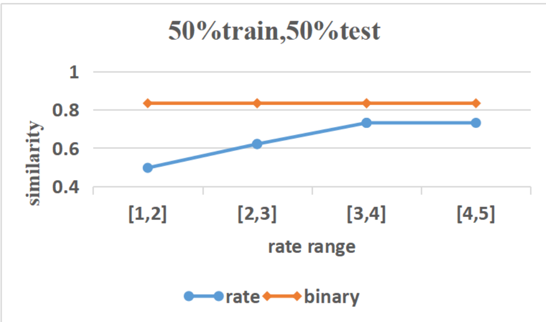

Figure 2: Divide 50% of the source data into training sets and 50% into test sets, using binary data and scoring data based on the experimental results in IV.

ISSN 2616-5775 Vol. 4, Issue 1: 94-104, DOI: 10.25236/AJCIS.2021.040115

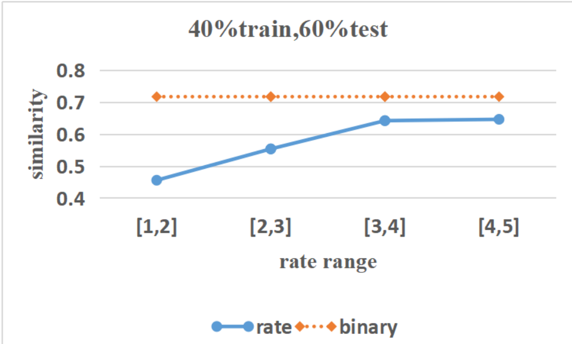

Figure 3: Divide 40% of the source data into training sets and 60% into test sets, using binary data and scoring data based on the experimental results in IV.

From Figure 1, 2, 3, we found that using binary method calculated the past and the future of film type of similarity degree is significantly higher than the different interval marking value calculated a lot, and score values at different points of similarity calculation are more likely the result of the difference, in general, the binary method is superior to score value method.

By analyzing the experimental results, we can draw a conclusion that the expression method using binary data is more consistent with users' interests than the rating method, and this consistency is crucial for collaborative filtering recommendation algorithm.

5.3.2 Comparison of Recommendation Effects between Binary Data and Rating Data

According to the experimental results in 1), preliminary verified the introduction of binary data in the recommendation algorithm instead of rate data can reflect user's interests more effectively, but it is only verification, there is no convincing, so based on the improved algorithm, under the unified standard of binary data and ratings on the experimental results, the influence of the experimental results by RMSE and F value to display, concrete as shown in Figure 4 and table 5:

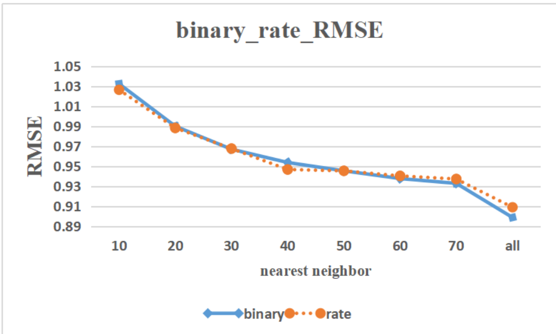

Figure 4: On the basis of the experiment in Part 3.3, the recommendation effect of binary data and rating data under different nearest neighbor conditions was compared, and RMSE was used as the evaluation index of recommendation effect.

Table 5: Binaries and rates (F values)

F values	Data type	Accuracy	Recall	F
1	Binaries	0.51603	0.47435	0.49431
2	Rates	0.51412	0.45925	0.48514

According to the data in Figure 4, as the nearest neighbor increases, the RMSE value using binary data in the algorithm becomes smaller and smaller, and ultimately lower than the result obtained by using rate data in the algorithm. The data in Table 5 shows the most important index data for the effect comparison of the recommendation algorithm. It can be seen that the binary data used in the algorithm is higher in accuracy, recall and the harmonic mean value (F) of accuracy and recall than the rate data. Combined with the experimental results of V) C) 1), it can be concluded that the use of binary data can not only locate users' interests more accurately, but also bring some optimization to the effect of recommendation algorithm.

ISSN 2616-5775 Vol. 4, Issue 1: 94-104, DOI: 10.25236/AJCIS.2021.040115

5.3.3 Comparison of Recommendation Effect between Improved Algorithm and Classical Algorithm

5.3.3.1 RMSE Comparison

In this paper, the improved algorithm for dynamic extraction of effective dense sub-sequences of project information and analysis of users' potential interest is respectively compared with classical algorithms such as SVD in the accuracy of recommendation results using RMSE index. The comparison results are shown in Figure 5:

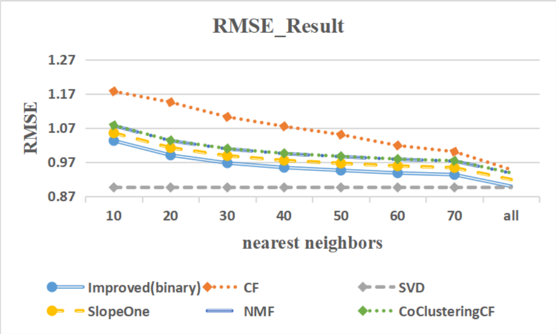

Figure 5: In the case of different nearest neighbor Numbers, the recommendation effect of improved binary algorithm is compared with that of classical algorithms.

Can be seen from the Figure 5 the proposed improved algorithm compared with other classic algorithms, in the case of different nearest neighbor, recommend the effect of accuracy is high, because the SVD algorithm is based on the matrix decomposition, so can't choose the nearest neighbor for its users, can only select all data to calculate, finally in the case of all users involved in the algorithm, the proposed improved algorithm under the evaluation standard for RMSE than SVD algorithm is more accurate.

5.3.3.2 F Value Comparison

The improved algorithm proposed in this paper compares the accuracy of the recommendation results with those of classical algorithms such as SVD by using f-value index. The comparison results are shown in Table 6:

Table 6: Recommended effect (F value)

F value				
	Name	Accuracy	Recall	F
1	Improved(binary)	0.51603	0.47435	0.49431
2	CF	0.49015	0.29588	0.36901
3	SVD	0.47492	0.48144	0.47816
4	SlopeOne	0.50447	0.47683	0.49033
5	NMF	0.49153	0.4624	0.4766
6	CoClusteringCF	0.4988	0.53457	0.46486

According to the value of F given in Table 6, the accuracy of the recommendation algorithm as an important evaluation index is higher than that of other algorithms, and the harmonic average value (F value) of accuracy and recall rate is also the highest among the algorithms. Combined with RMSE evaluation criteria in V) C) 3) a), the proposed algorithm is obviously better than several classical algorithms in terms of recommendation effect, and the recommendation accuracy has been significantly improved.

6. Conclusion

In this paper, the traditional collaborative filtering algorithm is improved by extracting efficient dense sub-sequences of project information and generating effective data. On this basis, the influence of user ratings on user interest is studied. The improved collaborative filtering algorithm can not only locate users' interest characteristics more accurately, but also alleviate the data sparsity problem of the collaborative filtering algorithm to some extent. From the experimental data and the results show that the improved algorithm has better recommendation effect, but in the improved algorithm, using the user whether the score(binary data) compared to rate values give recommendation algorithm optimization is not obvious, so the follow-up research will focus on how to use the user whether the rate(binary data) can be recommendation algorithm optimization effect maximization and explore more suitable for the research direction of similarity algorithm.

References

[1] Marken A. (2006) The Long Tail Why the Future of Business is Selling Less of More[J]. Hyperion, 24(3):274-276.
[2] Zhang S, Yao L, Sun A X, (2018) Tay Y.Deep learning based recommender system[J]. CM Computing Surveys, 1(1):1-35.
[3] Hong L, Ren Q H, Liang S X. (2016) Comparative Study on Information Service Quality of Recommendation System of Domestic E-commerce Websites -- A case study of Taobao,JINGdong and Amazon[J]. Library and Information Work, 60(23):97-110.
[4] Ma X T. (2016) Research on library Personalized Service Recommendation System based on Situational Big data[J]. Modern Intelligence, 36(4):90-94.
[5] Amatriain X, Pujol J M, Tintarev N. (2009) Rate it again:increasing recommendation accuracy by user re-rating[C]. ACM Conference on Recommender Systems. ACM, pp. 173-180.
[6] Young O K. (2008) Improving top-N recommendation techniques using rating variance[C]. In Proceedings of the Conference on Recommender Systems (RecSys'08), ACM Press, New York. NY. pp. 307–310.
[7] Jugovac M, Jannach D. (2017) Interacting with Recommenders—Overview and Research Directions[J]. ACM Transactions on Interactive Intelligent Systems, 7(3):1-46.
[8] Xing C X, Gao F R, Zhan S N, Zhou L Z. (2007) A collaborative filtering recommendation algorithm that ADAPTS to user interest changes[J]. Computer Research and Development, pp. 296-301.
[9] Liu K, Shi X, Kumar A. (2016) Temporal learning and sequence modeling for a job recommender system[C]. Proceedings of the Recommender Systems Challenge, ACM, pp. 1-4.
[10] Ding S H, Ji D H, Wang L L. (2015) Collaborative filtering recommendation algorithm based on user attributes and ratings[J]. Computer Engineering and Design, 36(2):487-491+497.
[11] Mao Y Y, Liu J H, Hu R, Tang M D, Shi M. (2017) A collaborative filtering recommendation algorithm for Web services using Sigmoid function[J]. Computer Science and Exploration, 11(2):314-322.
[12] Qian R, Wu Y, Kong G Q. (2018) Research on collaborative filtering algorithm based on fusion sparsity weighting[J]. Computer Technology and Development, 28(7):21-24+42.
[13] Yu H, Li Z Y. (2010) Collaborative filtering recommendation algorithm based on forgetting curve[J]. Journal of Nanjing University (Natural Science Edition), 46(5):520-527.
[14] Jia W Y, Li S Q, Li X Y, Liu B. (2018)A collaborative filtering recommendation algorithm based on discrete variables and user interest proximity[J]. Computer Engineering, 44(1):226-232+237.
[15] Benjamin M M, Richard S Z. (2009) Collaborative prediction and ranking with non-random missing data[C]. In Proceedings of the 3rd ACM Conference on Recommender Systems (RecSys'09). ACM Press, New York. NY. pp. 5-12.
[16] Koren Y. (2010) Factor in the Neighbors: Scalable and Accurate Collaborative Filtering[J]. Acm Transactions on Knowledge Discovery from Data, 4(1):1-24.
[17] Guo H Y, Liu G S, Su B, Meng K. (2016)Collaborative filtering recommendation algorithm combining community structure and interest clustering[J]. Computer research and development, 53(8):1664-1672.
[18] Liu Q H. (2016) A hybrid recommendation study based on community of interest and trusted neighbors[J]. Information science, 34(2):65-69.
[19] Yu B, Yang H L, Len M. (2018) Recommendation algorithm based on user interest model[J]. Computer System Application, 27(9):182-187.
[20] He M, Xiao R, Liu W S, Sun W. (2017) Collaborative filtering recommendation algorithm combining category information and user interest[J]. Computer Science, 44(8):230-235+269.
[21] Zhao W T, Wang C C, Cheng Y F, Meng L J, Zhao H H. (2016) Collaborative filtering algorithm based on user multiple attributes and interests[J]. Computer Application Research, 33(12):3630-3633+3653.
[22] Dong L Y, Wang Y Q, He J N, Sun M H, Li Y L. (2017) Collaborative filtering recommendation algorithm based on time attenuation[J]. Journal of Jilin University (Engineering Science), 47(4):1268-1272.
[23] Ren L. (2015)A collaborative recommendation algorithm combined with scoring time characteristics[J]. Computer applications and software, 32(5):112-115.

ISSN 2616-5775 Vol. 4, Issue 1: 94-104, DOI: 10.25236/AJCIS.2021.040115

[24] Wu Y T, Zhang X M, Wang X M, Li H. (2016)Collaborative filtering algorithm based on user fuzzy similarity[J]. Journal of communication, 37(1):198-206.
[25] Liu J, Wu W Q, Li X, Liu Y L, (2017) Wang J F.Collaborative filtering algorithm based on user interests and project attributes[J]. Computer Applications and Software, 34(5):33-37.