Research on Countermeasures of Network Collaborative Filtering Based on Genetic Algorithm

Xuechun Yang[1], Xing Ma[2], Shuai Zheng[1], Yongmei Bao[1]*

[1]Liaoning Nationalnormal College, Fuxin, 123000, Liaoning, China

[2]Liaoning University, Shenyang, 111000, Liaoning, China

Keywords: Genetic algorithm; Network CF; Countermeasure

Abstract: CF(Collaborative Filtering) also faces some problems. Such as cold start problem, scalability problem and data sparsity problem. The cold start problem is divided into system cold start, user cold start and project cold start. By using GA, the differentiation of user similarity is improved, which is beneficial for rating prediction and recommendation. Different from the classic CF algorithm, we use the constructed GA to predict the vacancy value in the scoring matrix before calculating the similarity. The crossover operation in GA selects the parent chromosome based on the crossover probability, and generates new chromosomes through crossover calculation. This paper attempts to study the countermeasures of network CF problem based on GA(Genetic Algorithm). The experimental results show that when the data is extremely sparse, the Mae of various algorithms will increase, but the algorithm in this paper has the minimum Mae value, which keeps around 0.79 on average, reflecting that the algorithm in this paper can better adapt to the data with different sparsity. Through the matrix filled with GA, the similarity of users is calculated, and the nearest neighbor set of the target user is found. In the process of crossover and mutation. The evolution of individuals is easy to produce infeasible individuals. After we get the nearest neighbor set of the target user, we can predict the user's score of unrated items according to the evaluation of the user's nearest neighbors.

1. Introduction

With the explosive growth of network information, users are facing an increasingly serious problem of information overload. In order to solve the problem of excessive information, users can quickly find the information they need from a large amount of data. Memory based co filtering includes three main steps: data normalization, neighbor selection, and determining the weight of interpolation. Currently, user based and project based neighbor selection methods have been widely studied. CF also faces some problems. For example, cold start issues, scalability issues, and data sparsity issues. Among them, data sparsity is a key factor affecting the recommendation quality of CF systems, which often leads to low recommendation accuracy[1]. In this paper, the network CF problem is studied based on GA. GA takes fitness function as its evolutionary goal. It can only evolve in the direction of increasing the value of the fitness function. Therefore, the fitness function and the objective function should be properly converted[2]. The crossover operation in GA selects the parent chromosome based on the crossover probability, and generates new chromosomes through crossover calculation. Firstly, two chromosomes are selected from the parent generation, and the number of swapping weights between chromosomes is determined by the probability of swapping. GA has better scalability and correctness, so this algorithm focuses on project-based methods, and the key step is correlation calculation. The mixed strategy of fitness ratio method and elite retention is used to determine the selection probability. First, the individual with the highest fitness in the population is directly copied to the next generation population. Then select the remaining population, and the selection proportion is based on the size of the fitness function. Calculate similarity for users through the matrix filled with GA, this algorithm can discover more potential user neighbors with dependency relationships; On the other hand, filling the user rating matrix in stages based on the reliability of user neighbors partially compensates for the problem of

  DOI: 10.25236/icmmct.2023.028

data sparsity[3-4].

2. Existing problems of collaborative filtering algorithm

CF algorithm is the most successful personalized recommendation technology at present. It is an algorithm to find the nearest neighbor set of the target user according to the comparison between the behavior information of the target user and the behavior of other users, and make recommendation predictions for the target user according to the interests of the similar neighbors of the target user to help them make decisions [5]. The shortcomings of existing CF algorithms are sparsity, scalability and cold start. In practice, the number of users and item days is very large. In this case, the scoring matrix will be extremely sparse, which is commonly referred to as sparsity, and has a negative impact on the CF algorithm. In large-scale e-commerce websites, with the increasing number of users and products, the dimension of user rating matrix becomes very high, and the number of products scored by users only accounts for a small part, so sparsity becomes the core factor affecting the recommendation effect of recommendation system [6-7]. CF algorithm is based on users' CF recommendation. By analyzing the scoring data of users' history, the nearest neighbor of the target user is found, and then the unknown interest of the target user is predicted by the scoring of the nearest neighbor. Because of this problem, the similarity between two users is very likely to be zero. This situation is called neighbor transmission loss. For example, if there is a high correlation between users and users, and there is also a high correlation between user mouth and user bin, then there is not necessarily a high correlation between users [8]. There are many methods to alleviate the sparseness of scoring matrix, although all of them alleviate the sparseness of data to some extent, but none of them are from the perspective of finding more potential dependent user neighbors and predicting the filling of scoring matrix in stages by decreasing reliability. They may have few common scores, and may even have a negative correlation because they have few scores.

3. Research on Countermeasures of Network Collaborative Filtering Problem Based on Genetic Algorithm

3.1. Correlation Calculation Based on Genetic Algorithm

Usually, because users only rate a portion of a large number of projects, most of the ratings are missing. In this case, if the user only rates a small portion, the Pearson correlation coefficient will overestimate the correlation between them. This article uses GA to predict missing scores to obtain a missing score matrix, and then uses the enhanced Pearson correlation method to calculate the correlation to form neighbors to obtain the predicted value of missing scores [9]. Therefore, the median of the rating interval can be used to represent the boundary point of interest. When the rating is above the median, it indicates liking, and when the rating is below the median, it indicates dislike. For a single user, the average rating of the user may better represent the boundary point of their interest tendency.

In the constructed user scoring matrix of the CF based recommendation system. Different from the classic CF algorithm, we use the constructed GA to predict the vacancy value in the scoring matrix before calculating the similarity. The crossover operation in GA selects the parent chromosome based on the crossover probability, and generates new chromosomes through crossover calculation. Firstly, two chromosomes are selected from the parent generation, and the number of swapping weights between chromosomes is determined by the probability of swapping. GA has better scalability and correctness, so this algorithm focuses on project-based methods, and the key step is correlation calculation[10]. Therefore, the fitness function and the objective function should be properly converted. If the network error in evolution is a positive non-zero number, the reciprocal of the objective function can be taken as the fitness function. Usually, Pearson correlation coefficients are used to provide a more reliable correlation formula based on parameters, which can be described as:

$$sim(i,j)=\frac{N_i \cap N_j}{N_i + N_j + \alpha} \cdot sim \tag{1}$$

Where: $N_i + N_j$ is the number of users who score on item i and item j at the same time; α is a parameter for adjusting correlation calculation.

Then the filled matrix is used to calculate the similarity of users and find the nearest neighbor set of target users. In the process of crossover and mutation. In this paper, the method of formula (1) is used to predict the target user i score P_{in} of unrated item a.

$$P_{in} = R_i + \sum_{m \in BU_i} sim(i,u) \tag{2}$$

$sim(i,u)$ represents the similarity between the target user i and the user u. R_i stands for the average score with i.

When the number of items jointly commented by two users is small, even if their scores are highly similar, it does not mean that they are necessarily similar, that is, the confidence of similarity is related to the number of items jointly evaluated by users. Table 1 shows the similarity between users calculated according to formula (2).

Table 1 Calculated Similarity

	User1	User2	User3	User4
User1	-	0.345	0.124	0.378
User2	0.315	-	0.258	0.357
User3	0.254	0.567	-	0.453
User4	0.287	0.361	0.497	-

It can be seen that equation (2) avoids the drawbacks of traditional similarity algorithms, improves the differentiation of user similarity, and is conducive to rating prediction and recommendation. In CF, users score items because they are interested in the implicit characteristics of these items, and there are some common characteristics among these items. The user's preference for an item is that the user has a high score for these items.

3.2. Experimental results and analysis

Sparsity directly affects the quality of this recommendation system, so it has attracted great attention from academic and application circles. For the cold start problem, a solution is to recommend users based on the similarity of users' attributes when the number of users' ratings is small, and with the increase of users' ratings, users' ratings are gradually used to make recommendations. How to integrate user attribute recommendation and user rating recommendation to achieve a smooth transition is a difficult problem. In order to test the cold start performance of the algorithm, the experiment randomly selects 25 users from the data set, deletes their related rating information, simulates the cold start scenario, and tests the cold start performance of the algorithm in this paper.

From the results in Figure 1, it can be seen that this method can fully utilize user information and to some extent alleviate the problem of cold start. In order to test the performance of the algorithm under different sparse data conditions, a portion of the scoring data was randomly removed from the dataset to artificially improve data sparsity. Then, the algorithm proposed in this paper was compared with traditional algorithms, and the results are shown in Figure 2.

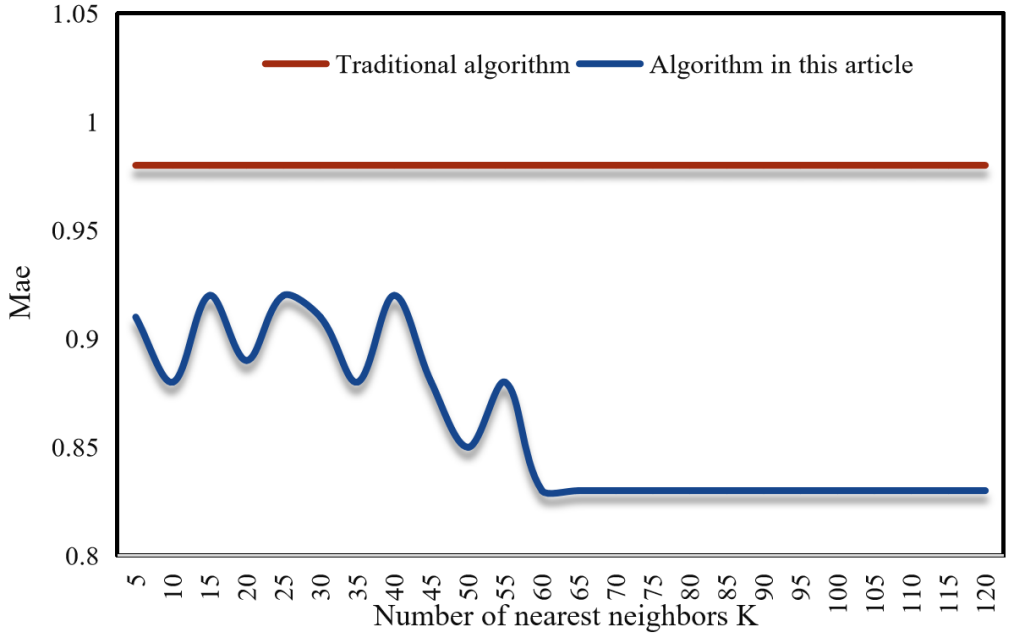

Figure 1 Algorithm performance under cold start conditions

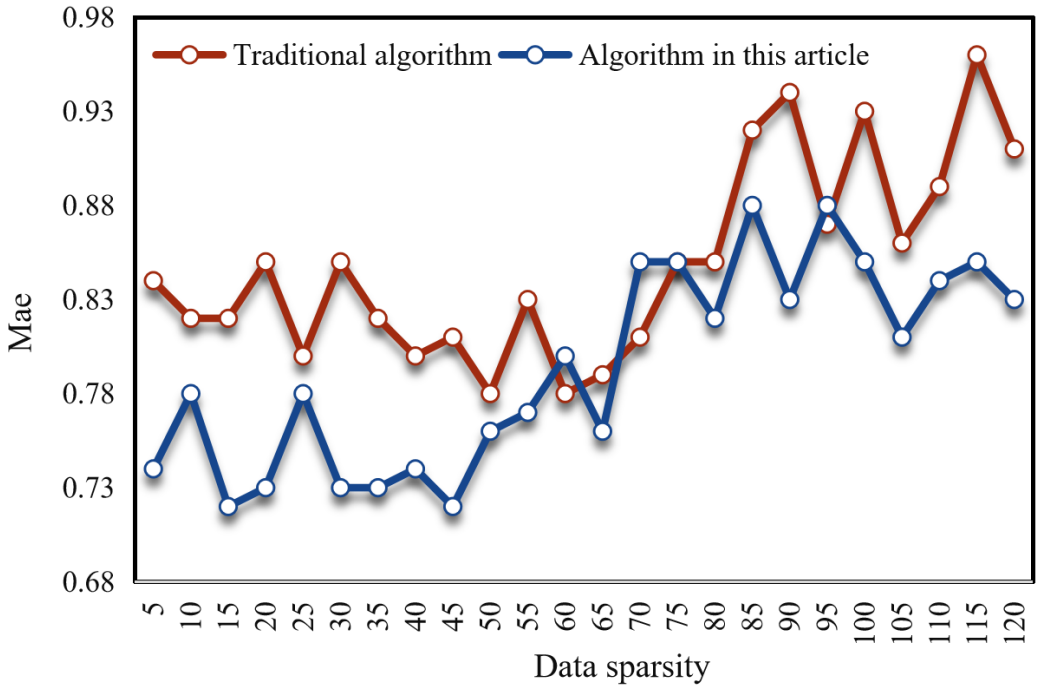

Figure 2 Performance of Different Algorithms under Sparsity

From Figure 2, it can be seen that when the data is extremely sparse, the Mae of various algorithms will increase. However, the algorithm in this paper has the minimum Mae value under various sparsities, maintaining an average of around 0.79, reflecting that the algorithm in this paper can better adapt to data with different sparsities.

Through the above experimental analysis, the median of the rating interval can be used to represent the boundary point of interest. When the rating is above the median, it indicates liking, and when the rating is below the median, it indicates dislike. For a single user, the average rating of the user may better represent the boundary point of their interest tendency when there are a large number of users and projects, all the default values are filled, and the amount of calculation to complete the recommendation is also relatively large, so it is suitable for small-scale databases. In addition, there will be some differences in users' scoring of projects that are not over-rated. This method uses unified values to fill in, without considering the differences of users' interests and

obliterating users' personality.

4. Conclusions

In this paper, we use GA to study the countermeasures of network CF problem, and effectively optimize the weights to solve the shortcomings of slow convergence of network speed and sensitivity of initial weights. By using GA, the differentiation of user similarity is improved, which is beneficial for rating prediction and recommendation. The article proposes a GA based approach to address the efficiency issues and low quality recommendation results when dealing with large datasets. This algorithm can discover more potential user neighbors with dependency relationships; On the other hand, filling the user rating matrix in stages based on the reliability of user neighbors partially compensates for the problem of data sparsity. We have achieved a smooth transition from similarity calculation in the constructed user scoring matrix of the CF based recommendation system. Different from the classic CF algorithm, we use the constructed GA to predict the vacancy value in the scoring matrix before calculating the similarity. The crossover operation in GA selects the parent chromosome based on the crossover probability, and generates new chromosomes through crossover calculation. Firstly, two chromosomes are selected from the parent generation, and the number of swapping weights between chromosomes is determined by the probability of swapping. The experimental results show that when the data is extremely sparse, the Mae of various algorithms will increase. However, the algorithm in this paper has the minimum Mae value under various sparsities, maintaining an average of around 0.79, reflecting that the algorithm in this paper can better adapt to data with different sparsities. The next step will be to study the number of steps for neighboring users to fill in based on their influence, in order to more reliably compensate for the sparsity of the rating matrix data.

Acknowledgements

Supported by Liaoning Provincial Public Welfare Research Fund for Science (Soft Science Research Program) in 2022["Current Situation and Research on the Construction of Scientific and Technological Talents in Colleges and Universities for Nationalities"] (No.: 2022JH4/10100030)

References

[1] Houshmand-Nanehkaran F, Lajevardi S M, Mahlouji-Bidgholi M. Optimization of fuzzy similarity by genetic algorithm in userbased CF recommender systems[J]. Expert systems: The international journal of knowledge engineering, 2022, 37(4):39-55.

[2] Alhijawi B, Kilani Y. A CF recommender system using genetic algorithm[J]. Information Processing & Management, 2020, 57(6):102310-102348.

[3] Laishram A. Discovery of user-item subgroups via genetic algorithm for effective prediction of ratings in CF[J]. Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies, 2019, 49(11):61-78.

[4] Chu Z, Han Y, Zhao K. Botnet vulnerability intelligence clustering classification mining and countermeasure algorithm based on machine learning[J]. IEEE Access, 2019, 112(99):24-51.

[5] PengXU, XuLIU, HuafengCAO, et al. An efficient energy aware virtual network migration based on genetic algorithm[J]. Frontiers of Computer Science in China: English Version, 2019, 13(2):3-14.

[6] Wang Y, Xu Y, Dan C. Research on CF Algorithm Based on SGA-RBF[J]. Statistics & Decision, 2019, 44(14):51-66.

[7] Pham T, Sanner J, Morin C, et al. Virtual network function–forwarding graph embedding: A genetic algorithm approach[J]. International Journal of Communication Systems, 2020, 33(10):18-

27.

[8] Nurika, Okta, Hassan, et al. A study of fluctuations and confidence of implementation in genetic algorithm optimized network in data centre[J]. Intelligent decision technologies: An international journal, 2018, 12(1):25-37.

[9] Cui X, Yang J, Li J, et al. Improved Genetic Algorithm to Optimize the Wi-Fi Indoor Positioning Based on Artificial Neural Network[J]. IEEE Access, 2020, 52(99):11-32.

[10] Tm A, Amb A, Re B, et al. Efficient clustering in CF recommender system: Hybrid method based on genetic algorithm and gravitational emulation local search algorithm[J]. Genomics, 2019, 111(6):1902-1912.