An improved user-based collaborative filtering algorithm

Zhiquan Zou
College of information science and engineering
Shandong Agricultural University
Tai an, China

Suming Zhang
College of resources and environment
Shandong Agricultural University
Tai an, China

Zhijun Wang*
College of information science and engineering
Shandong Agricultural University
Tai an, China

Shuhan Cheng*
College of information science and engineering
Shandong Agricultural University
Tai an, China

***Abstract*—The collaborative filtering algorithm[1] proposed by Grouplens[2] is one of the most commonly used methods for personalized recommendation in recommendation systems[3][4][5][6],and the core component of User-based collaborative filtering is the similarity measure. The traditional user similarity measurement method does not consider the influence of factors such as frequent user interest transfer and content popularity degree difference on the accuracy of the algorithm, and the existing improvement strategies cannot comprehensively consider these two factors. Based on the traditional similarity algorithm, this paper introduces influential factors such as user interest decline over time and content popularity, so as to improve the existing user similarity algorithm and to compare the actual data to prove the improved algorithm.**

Keywords-component; Adjusted Cosine similarity; Pearson Correlation; Recommendation; Similarity measure; User-based collaborative filtering;

I. INTRODUCTION

With the advent of the Internet age, the process of human social information is once again accelerating, and users' daily online behavior will generate a large amount of data. With the development of computer technology, it has been able to efficiently process this user behavior data. For enterprises, this part of the data directly drives the development of an enterprise or product, and the use of data has gradually become a key factor in the core competition among enterprises. The user evaluation data is an important part of the user behavior data. The user recommendation service formed on the basis of this data has become the main means to improve the user holding rate.

The recommendation technologies currently used in the recommendation system mainly include Association Rules[7], Content-based Recommendation[8], Collaborative Filtering, and Hybrid Approach[9]. Among them, collaborative filtering algorithm proposed by GroupsLens is the most widely used one[10][11].Moreover, the nearest neighbor collaborative filtering recommendation is currently the most successful recommendation technique[12] .

With the increase of product operating time, the content system is gradually perfecting, the number of content is increasing sharply, and the user activity is climbing. As a result, the user evaluation cycle becomes shorter, the interest transfer is more frequent, and the content popularity is significantly different. The effect of these factors on the accuracy of user similarity measure in collaborative filtering algorithm is becoming more and more obvious. In order to solve the above problems, this paper introduces the time decay rule of user interest and the analysis strategy of content popularity to improve the existing user similarity measurement algorithm, so as to improve the accuracy of user similarity analysis, so as to reach the goal of improving the quality of the push. Finally, a comparison experiment was conducted on the actual data set, and the test results were analyzed and evaluated by Mean Absolute Error(MAE)[13]. The results show that the improved calculation method is significantly better than the traditional recommendation strategy and can more accurately calculate the user similarity. Achieve higher quality recommendations.

II. TRADITIONAL SIMILARITY ALGORITHM

The traditional user-based collaborative filtering algorithm recommends to the target user according to other user's preferences. It first finds a group of neighboring users that have the same preference as the target user, then analyzes the neighboring user and recommends neighboring users' favorite items to the target user [14][15][16][17].

TABLE I. User rating matrix indicates that an m*n matrix can simply represent the data model of the user rating, m rows indicate that there are m users, n columns indicate that there are n columns of items, $R_{i,j}$ indicates that the i-th user evaluates the score for the j-th item. The time matrix structure corresponding to the user rating is the same as the following scoring matrix, and the matrix element is the time stamp of the rating. In this paper, the traditional user similarity algorithm and the improved user similarity algorithm are based on these two matrices for similarity measurement.

978-1-5386-8097-1/18/$31.00 ©2018 IEEE

TABLE I. USER RATING MATRIX

User \ Item	Item-1	Item-2	...	Item-k	...	Item-n
User-1	$R_{1,1}$	$R_{1,2}$	...	$R_{1,k}$	...	$R_{1,n}$
User-2	$R_{2,1}$	$R_{2,2}$	...	$R_{2,k}$	...	$R_{2,n}$
...	...	...	...	...	...	...
User-i	$R_{i,1}$	$R_{i,2}$	...	$R_{i,k}$	...	$R_{i,n}$
...	...	...	...	...	...	...
User-m	$R_{m,1}$	$R_{m,2}$	...	$R_{m,k}$	...	$R_{m,n}$

A. AdjustedCosine similarity

The adjusted cosine similarity algorithm is based on the cosine similarity[18]. It is insensitive to the numerical calculation of the cosine similarity and does not consider the differences in the dimension of the user. The adjusted cosine is obtained by subtracting the mean value. The similarity eliminates the differences in the dimensions of user dimensions and provides a more complete calculation strategy for similarity calculation. (1) denotes the similarity function calculated using the adjusted cosine similarity. $R_{i,k}$, $R_{j,k}$ represent the user i's , j's score for the item k respectively, $\bar{R}_i$,$\bar{R}_j$ is the user's average score for item k in the $item_{ij}$ space.

$$\text{adjustedCosine}(i,j) = \frac{\sum_{k\in item_{ij}}(R_{i,k}\text{-}\bar{R}_i)(R_{j,k}\text{-}\bar{R}_j)}{\sqrt{\sum_{k\in item_{ij}}(R_{i,k}\text{-}\bar{R}_i)^2}\sqrt{\sum_{k\in item_{ij}}(R_{j,k}\text{-}\bar{R}_j)^2}} \quad (1)$$

B. Pearson Correlation

The pearson correlation coefficient is also a current excellent similarity algorithm used to measure the linear correlation between two variables. As shown in (2), pearson(i,j) represents a function that uses the Pearson correlation coefficient to perform a similarity calculation, and $R_{i,k}$, $R_{j,k}$ represent user i,j scores for item k, respectively. $\bar{R}_i$,$\bar{R}_j$ are the user's average ratings for item k in the $item_{ij}$ space.

$$\text{pearson}(i,j) = \frac{\sum_{k\in item_{ij}}(R_{i,k}\text{-}\bar{R}_i)(R_{j,k}\text{-}\bar{R}_j)}{\sqrt{\sum_{k\in item_i}(R_{i,k}\text{-}\bar{R}_i)^2}\sqrt{\sum_{k\in item_j}(R_{j,k}\text{-}\bar{R}_j)^2}} \quad (2)$$

From the formula point of view, the Pearson correlation coefficient and the adjusted cosine similarity algorithm are similar. The covariance of two variables is divided by the corresponding standard deviation of each vector. Adjusted cosine similarity algorithm considers the mean value of each user that has graded, for and the Pearson correlation coefficient considers each $item_i$ that has been graded by user . The average of the scored points, the average length of the user set to which the average calculation belongs differs.

III. ALGORITHM IMPROVEMENT STRATEGY

Due to the increasingly active user rating behavior, the richness of Internet rich media resources, and the improvement of the evaluation system, users tend to change rapidly, and the rate of change of user evaluation data is relatively high. It is necessary to consider factors such as the factors of interest decline over time and the popularity of the project. Into the specific similarity calculation, to improve the accuracy of the similarity calculation.

A. Add Time Decrease Impact Factor

In terms of statistics, a person's preference for a certain type of thing gradually declines with time. The closer the time of comment or rating to the current time is, the more likely it is to reflect the current tendency of a user, and it should have a higher weight. Has a lower weight. According to Newton's law of cooling, when there is a temperature difference between the surface and the surroundings, the heat lost per unit area per unit time is directly proportional to the temperature difference, and the proportional coefficient is the heat transfer coefficient.

$$\frac{dT(t)}{dt} = -k(T(t)\text{-}H) \quad (3)$$

As shown in (3), T(t) represents the current temperature of the object, H is the surrounding temperature, and k is the specific decay coefficient.

When the ambient temperature H is 0, the following solution is obtained:

$$T(t) = T(t_0)e^{-kt} \quad (4)$$

According to (4), suppose that the user's score on an item is the "initial temperature" of the item, and the actual current value of the item for the user is calculated according to Newton's cooling law. As shown in (5), $f_{time}(t)$ represents the user's current favorite heat, f0 represents the favorite heat at the time of evaluation, t is the current and current time difference, and k is the specific decay coefficient.

$$f_{time}(t) = f_0 \cdot e^{-kt} \quad (5)$$

B. Add item popularity influence factor

According to statistics, in a user's evaluation, browsing, or purchase record, the higher the purchase or browsing frequency or rating of a product or content, the better the importance of such goods or content to the user. The user's personalized preferences should be given a higher weight. However, if the product or content in the entire data set for the majority of users, the higher the frequency of purchase, browsing, or rating, it indicates that this is a more popular item, can not accurately describe the user's personalized preferences, based on a more unpopular items are more likely to represent and describe user preferences. The algorithm idea is derived from term frequency–inverse document frequency(TF-IDF). The weighting techniques used for information retrieval and information mining are used to calculate the popularity of items in user similarity calculations, and a scoring strategy is applied to weighted processing. Based on the above principle, you can get (6).

$$f_{hot} = \frac{R_{i,k}}{\sum R_i} \cdot log \frac{U_{sum}}{U_R+1} \quad (6)$$

In (6), $R_{i,k}$ is the evaluation score of the user, U_{sum} is the sum of the number of users, and UR is the number of users who have evaluated this item.

C. Improved user similarity algorithm

Comprehensive (5), (6), The final influencing factor is shown in (7).

$$f_{weight} = f_{time} \cdot f_{hot} \quad (7)$$

1) The f_{weight} function is used to perform correction processing on the data in the scoring matrix used to calculate user similarity, and the effect of time decay is eliminated by the ftime function, so that the scoring matrix can more reflect the user's current preferences and tendencies. By the fhot function, penalizing the hot item score in the user's rating is a user's rating that represents the user's personalized preferences much better.

2) Using the modified scoring matrix, the target user's similarity calculation is performed by the adjusted cosine similarity and the Pearson correlation coefficient to obtain the nearest neighbor set.

D. Generate recommendations

Through the improved user similarity algorithm, after obtaining the nearest neighbor set of the target user who has watched the movie, the corresponding scoring prediction is performed with the actual scoring matrix to complete the recommendation. Supposing the set of nearest neighbors of user u is denoted by NBS_u, Then user U's prediction score P $P_{u,i}$ for item i can be obtained by user U's scoring of the items in the nearest neighbor collection NBS_u, and sim(u,n) is the similarity formula used specifically. The calculation method is as (8)[19].

$$p_{u,i} = \overline{R_u} + \frac{\sum_{n \in NBS_u} sim(u,n)(R_{n,i} - \overline{R_n})}{\sum_{n \in NBS_u}(|sim(u,n)|)} \quad (8)$$

IV. Empirical Study

A. Experimental data set description

Dataset is MovieLens provided by Grouplens, the data size is 500000 user evaluation records, the number of users is 943, the number of movies is 1650. The scores are 0 to 5.According to the experiment in this paper, the corresponding basic data set and test data set are allocated in a ratio of 80% and 20%.

In the time-dependent calculation process, because the current time and the scoring time differ greatly, in order to ensure correctness of the experimental calculation, the timestamp closest to the current time in the data set is used, and the next day of the timestamp is taken as the current time.

B. Evaluation criteria

This paper improves the user similarity algorithm based on user similarity in collaborative filtering algorithm. The final measure is the difference between the predicted score after similarity calculation and the actual score of users.

In this paper, MAE is used as a specific measurement standard. The smaller the MAE value, the closer to the actual score, the more accurate the prediction. If the MAE value is obtained according to the calculation method in this paper is smaller than that obtained by traditional similarity calculation, then the improvement of the similarity calculation method in this paper is effective and the final prediction of recommendation service is improved.

Let the user's predicted score set be $\{p_1,p_2,p_3,p_4,p_5...\}$ and the user's actual score set is $\{q_1,q_2,q_3,q_4...\}$.Then MAE is defined as (9).

$$MAE = \frac{\sum_1^N (|p_i - q_i|)}{N} \quad (9)$$

C. Experimental results

In the following experimental results, the vertical coordinate is MAE value, the vertical coordinate interval is 0.05, the horizontal coordinate is the number of nearest neighbors, and the horizontal coordinate interval is 2.

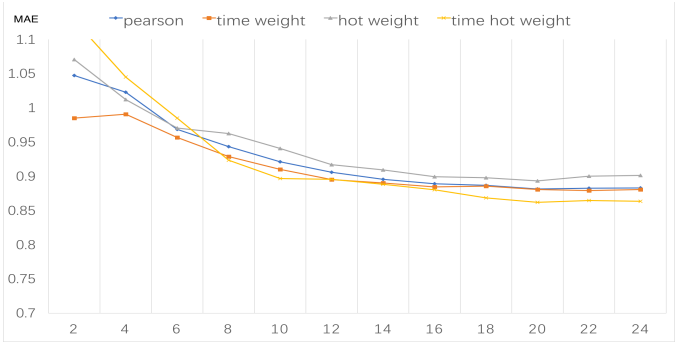

Figure 1. Comparing in MAE for pearson

1) Comparing in MAE for pearson: As shown in Figure 1, Pearson's broken line represents the MAE result of the traditional calculation of user similarity using Pearson correlation coefficient. The time weight broken line represents the MAE result of user similarity calculation using Pearson correlation coefficient after adding time decay factor to get the rating prediction. Hot weight broken line represents the MAE result of rating prediction using Pearson correlation coefficient after adding the influence factor of project popularity. Hot time weight broken line represents the MAE result of user similarity calculation using Pearson correlation coefficient to obtain the rating prediction by adding time recession impact factor and project popularity impact factor at the same time. It can be seen when using Pearson correlation coefficient for user correlation calculation only add time recession all impact factors of MAE value less than traditional Pearson correlation coefficient of similarity calculation of MAE

value, add time recession impact factor for the calculation of Pearson correlation coefficient to improve the effect. When both factors are added, when the number of nearest neighbors reaches 8, the MAE value is significantly lower than the traditional Pearson correlation coefficient calculation method, and the algorithm is improved significantly.

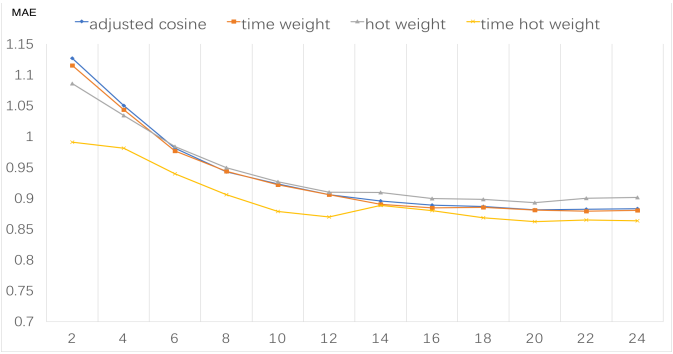

Figure 2. Comparing in mae for adjusted cosine

2) Comparing in mae for adjusted cosine: For the calculation of adjusted cosine similarity, the improvement effect of adding time recession influencing factor and project popularity influencing factor is more obvious. As shown in Figure 2, the adjusted cosine broken line represents the MAE result of traditional evaluation prediction using adjusted cosine similarity calculation. The time weight polyline represents the MAE result of the rating prediction calculated by the adjusted cosine similarity after adding the time decay factor. The hot weight polyline represents the MAE result of rating prediction based on the adjusted cosine similarity calculation of the project popularity factors. The hot time weight polyline represents the MAE result of user similarity calculation using adjusted cosine to obtain the score prediction. Adjusted cosine similarity add time recession after impact factor and project popularity impact factor, the nearest neighbor number is 2 to 24 situations were significantly lower than the traditional fixed cosine similarity between calculated MAE value, especially between nearest neighbor number is 2 to 14, MAE improvement effect is most obvious.

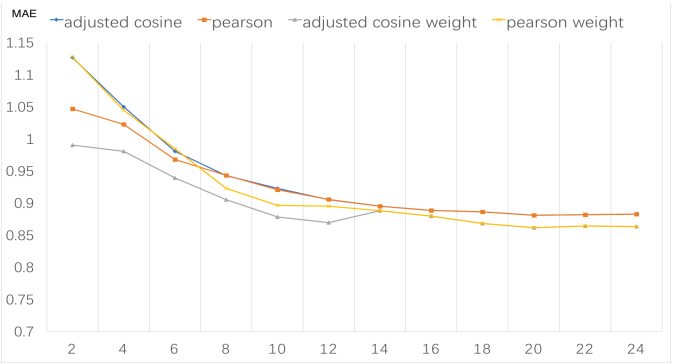

Figure 3. Comparing in mae for two algorisms

3) Comparing in mae for two algorisms: As shown in Figure 3, the adjusted cosine broken line represents the MAE result of traditional evaluation prediction using adjusted cosine similarity calculation. Pearson's broken line represents the MAE result of the traditional calculation of user similarity using Pearson correlation coefficient. Adjusted cosine weight broken line represents the MAE result of user similarity calculation using adjusted cosine weight to obtain score prediction by adding time recession impact factor and project popularity impact factor at the same time. Pearson weight broken line represents the MAE result of user similarity calculation using Pearson correlation coefficient to obtain the rating prediction by adding time recession impact factor and project popularity impact factor at the same time. Adjusted cosine similarity calculated after add impact factor score predicts all MAE value and obvious less than traditional way, and improve the effect is better than that of add after impact factor calculated using Pearson correlation coefficient. From the experimental results it is concluded that, based on the score predicts to recommend recommendation service, add time recession factors and project popularity can obviously improve the accuracy of the prediction and improving the quality of the corresponding recommendations, to the improvement of the recommendation system effect is remarkable.

D. Analysis of results

The biggest difference between the method proposed in this paper and the traditional method is that the user's similarity calculation process adds time impact factors that are closely related to people's subjective interests, and the popularity of the project.

Experiments conducted on different number of nearest neighbors found that the MAE in the score prediction was significantly reduced. In the experimental process, it was found that when only the factors of the time recession or the popularity of the project were added, the effect was not obvious or even under certain circumstances. Higher than the MAE in traditional calculations, but simultaneously adding these two influencing factors, the Pearson correlation coefficient and the adjusted cosine similarity measure have been significantly improved, especially for the improvement of the adjusted cosine similarity is extremely obvious, the score prediction The MAE in time are from 2 to 24, and the interval is 2, and the whole is smaller than the two traditional methods and the improved Pearson correlation coefficients are used to measure the similarity.

The experimental results and analysis show that the MAE of the score prediction becomes smaller and the score of the prediction is significantly lower when the subjective interest points introduced in the similarity measure are reduced over time and the project's unpopularity is more representative of a person's tendency. More accurate, it can significantly improve the recommendation quality of the recommendation service based on the basic score prediction.

V. CONCLUSION

This paper analyzes the collaborative filtering algorithm in the recommendation system and finds that when the total

duration of the user is used and the number of items increases, the user's interest changes with time and the difference in the popularity of the project becomes larger. When the user similarity is calculated, the influence factor becomes Can not ignore, need to introduce these influencing factors to improve the calculation of the traditional similarity measure. By analogy to Newton's cooling law in physics and the TF-IDF statistical strategy in text statistics, the calculation function corresponding to the time-decay function and the popularity degree in the recommendation scene is derived, thereby improving the existing user similarity measurement function. Control experiments on the MoiveLens dataset, test and analyze the improvement of the algorithm after adding these two influencing factors. The results show that the time decay factors and the popularity of the project added in the similarity measure can significantly reduce the MAE of the score, thus significantly improving the recommendation quality.

REFERENCES

[1] Li C. Research on the Bottleneck Problems of Collaborative Filtering in Ecommerce Recommender Systems. Ph. D Dissertation. Hefei, China: Hefei University of Technology, 2009.

[2] Konstan JA, Miller BN, Maltz D, Herlocker JL, Gordon LR, Riedl J. Grouplens: Applying collaborative filtering to usenet news. Communications of the ACM, 1997, 40(3): 77-87.

[3] Resnick P, Varian H R. Recommender Systems. Communications of the ACM, 1997, 40(3) : 56 – 58.

[4] Zenebe A, Norcio A F. Representation, Similarity Measures and Aggregation Methods Using Fuzzy Sets for Content-Based Recommender Systems. Fuzzy Sets and Systems, 2009, 160(1) : 76 – 94.

[5] Schafer J B, Konstan J A, Riedl J. E-commerce Recommendation Applications. Data Mining and Knowledge Discovery, 2001, 5(1/ 2): 115-153.

[6] Adomavicius G, Tuzhilin A. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Trans on Knowledge and Data Engineering, 2005, 17(6): 734-749.

[7] Su X N, Yang J L, Deng S H, et al. Theory and Technology of Data Mining. Beijing , China: Scientific and Technical Documentation Press,2003.

[8] Baeza-Yates R , Ribeiro-Neto B. Modern Information Retrieval. New York, USA: ACM Press, 1997.

[9] Xu H L, Wu X, Li X D, et al. Comparison Study of Internet Recommendation System. Journal of Software, 2009, 20(2) : 350 - 362 .

[10] Mooney R J. and Roy L, Content-Based Book Recommending, In Proceedings of the fifth ACM conference on Digital libraries , Berkeley, CA, 1999, 195 - 204.

[11] Blacker K D , Lawrence S , Giles C. L. ,Discovering Relevant Scientific Literature on the Web, IEEE IntelligentSystems and their Applications, 2000, 15(02) : 42 - 47.

[12] Breese J, Hecherman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI'98). 1998. 43~52.

[13] Sarwar B, Karypis G, Konstan J, Riedl J. Item-Based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International World Wide Web Conference. 2001. 285~295.

[14] Jeong B, Lee J, Cho H. An Iterative Semiexplicit Rating Method for Building Collaborative Recommender Systems. Expert Systems with Applications, 2009, 36(3) : 6181 – 6186.

[15] KarypisG.Evaluation of Item-Based Top-N Recommendation Algorithms Proc of the 10th International Conference on Information and Knowledge Management. Atlanta, USA, 2001: 247 – 254.

[16] Liang C Y, Leng Y J, Wang Y S, et al. Research on Group Recommendation in Ecommerce Recommender Systems. Chinese Journal of Management Science, 2013, 21(3) : 153 - 158 (in Chinese)

[17] de Campos L M , Fernández-Luna J M , Huete J F , et al. Combining Content-Based and Collaborative Recommendations: A Hybrid Approach Based on Bayesian Networks. International Journal of Approximate Reasoning, 2010, 51(7) : 785 – 799.

[18] SHI Kansheng, LIU Haitao, BAI Yingcai, et al.Text Clustering Method with Improved Fitnes Function and Cosine Similarity Measure.Journal of University of Electronic Science and Technology of China, 2013, 42(4):621-624.

[19] Xiang Liang. Recommended practice system. Beijing,China: POSTS&TELECOM PRESS, 2 0 1 2 :4 1 -6 4.