Collaborative filtering algorithm based on optimized clustering and fusion of user attribute features

Wu Qingyang
China Automobile Data of Tianjin Co., Ltd
wuqingyang@catarc.ac.cn

Cheng Xu
China Automobile Data of Tianjin Co. Ltd.chengxu@catarc.ac.cn

Sun Erxin
China Automobile Data of Tianjin Co. Ltd.sunerxin@catarc.ac.cn

Deng Chengpeng
China Automobile Data of Tianjin Co. Ltd.dengchengpeng@catarc.ac.cn

ABSTRACT

Aiming at the problems of low recommendation quality, low recommendation efficiency, and cold startup in the collaborative filtering recommendation algorithm, a collaborative filtering recommendation algorithm based on optimized K-means clustering algorithm and user attribute features is proposed. According to the user attribute information, the optimized K-means clustering algorithm is used to cluster them; Multiple clusters are generated, and a novel similarity calculation model is formed by combining user attribute features in each cluster; Considering that users' interests will change dynamically with time, time factor is introduced into traditional scoring similarity; Through this model, the nearest neighbor of the target user is found, and the recommendation list is generated to realize the recommendation. The experimental results produced on the MovieLens datasets show that this algorithm can shorten the algorithm operation time and solve the cold start problem while improving the recommendation efficiency and recommendation accuracy.

CCS CONCEPTS

• **Information systems** → Information systems applications; • **Computing methodologies** → Modeling methodologies.

KEYWORDS

Collaborative filtering, Optimized clustering, User attribute features, time factor, similarity calculation

ACM Reference Format:
Wu Qingyang, Cheng Xu, Sun Erxin, and Deng Chengpeng. 2021. Collaborative filtering algorithm based on optimized clustering and fusion of user attribute features. In *2021 4th International Conference on Data Science and Information Technology (DSIT 2021), July 23–25, 2021, Shanghai, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3478905.3478931

DSIT 2021, July 23–25, 2021, Shanghai, China

ACM ISBN 978-1-4503-9024-8/21/07…$15.00
https://doi.org/10.1145/3478905.3478931

1 INTRODUCTION

With the rapid development of mobile technology, more and more people have smart phones, tablet computers and other intelligent terminals, which makes the speed of information production explosive growth, leading to the problem of information overload,when users search the information they are interested in, it will take a lot of time, but the results are often not satisfied with the users, so personalized recommendation technology comes into being.Personalized recommendation technology refers to the use of a user's point of interest to recommend interested content to users, which is an effective way to alleviate the problem of information overload.Among the personalized recommendation technologies[1], collaborative filtering recommendation technology[2] is the most mature and widely used technology,in simple terms, collaborative filtering is to predict the information that users are interested in based on the preferences of a certain group of users who have similar interests and recommend it to target users.Although the collaborative filtering recommendation algorithm has many advantages compared with other recommendation algorithms, there are still typical problems common to recommendation algorithms, such as recommendation efficiency, low recommendation quality, cold start, etc.

For the above problems, many scholars have improved the recommendation algorithm.In view of the scalability problem of traditional collaborative filtering recommendation technology, in the face of an increasing number of users and the rapid increase in data volume, the scalability problem of the algorithm is called an important factor restricting the recommendation system.[3] proposed a collaborative filtering algorithm based on traditional SVD, but this matrix decomposition algorithm has a certain cost.[4] proposed a new collaborative filtering user recommendation algorithm based on K-means clustering, which was improved on the basis of traditional K-means clustering algorithm, and solved the user clustering problem well.In reference [5], the K-means clustering model of bee colony is proposed to cluster users and generate multiple clusters. The target user searches for the nearest neighbor in the cluster and generates the recommendation list to realize the recommendation.In reference [6], an extended naive Bayes hybrid recommendation algorithm based on probability classification is proposed, [7-9] proposed CF algorithms for clustering based on the user's own unique information, these algorithms have effectively alleviated the user's cold start problem, thereby improving the response speed of the recommendation.

Table 1: Initial clustering center

age	gender	occupation
0	0	2
1	0	2
2	0	3
3	0	3

Inspired by the above methods, this paper proposes a collaborative filtering recommendation algorithm based on optimized K-means clustering algorithm and user attribute features,First, clustering is performed based on user attribute features, the clustering uses the optimized K-means algorithm, and then a new similarity calculation algorithm is used to generate recommendations for the target users based on the target users after clustering.The results of using the MovieLens data set show that the algorithm can effectively solve the cold start problem of new users while improving the performance of the recommendation algorithm.

2 K-MEANS CLUSTERING BASED ON USER ATTRIBUTE FEATURES

Clustering [10-11] refers to the process of classifying similar objects into multiple classes in a collection of physical or abstract objects. These objects are similar to each other in the same class and are different in different classes.However, the traditional CF algorithm is often based on the user's rating data of the product clustering, while ignoring some characteristic attributes specific to the user, so there will be an embarrassing situation of recommending some information that is not of interest to the target user. It greatly affects the accuracy of the recommendation.In reality, each user has its own personal features. This paper will cluster users according to their gender, age, occupation and other attribute features, and think that users with similar age, gender and occupation may have similar preferences and consumption behavior.

K-means is the most commonly used clustering algorithm.In k-means algorithm, the selection of initial clustering center has a certain impact on the clustering results. In order to reduce the impact of initial clustering center on clustering results, this paper optimizes the selection of initial clustering center when using k-means clustering algorithm.Age, gender and occupation are the unique features of users. In this paper, when K-means clustering is used for user attribute features, the initial clustering center is the selection of K value, and K is selected according to the unique features of users (age, gender and occupation). According to the division of age, this paper can be divided into four stages,That is, K is 4.

Firstly, the most frequent user groups with high activity attribute features are selected as the selection objects, because the selection of clustering center is based on the division of age groups. In this paper, the four age groups are represented by numbers 0, 1, 2 and 3 respectively, and then the gender attributes with high activity (male and female, represented by numbers 0 and 1) are selected in each age group, According to age and gender, the users with high occupation activity in each age group are selected. In this paper, 21 kinds of occupation are classified into four categories, which are represented by the numbers 0, 1, 2 and 3 respectively. In this way, the users in each age group are selected in turn, and the results are shown in Table 1

Select the initial clustering center based on attribute feature activity according to Table 1, and then select the user ID set belonging to the cluster center of each row according to the user preference activity,and then based on the user ID set of user preference activity, the user with high score activity is selected as the initial clustering center, and the final K value selection is shown in Figure 1.

The initial cluster centers selected according to Figure 1, K-means clustering of users using user features attribute data sets.,The specific steps of the algorithm are as follows:

Algorithm 1 Algorithm 1

inputuser features information table、scoring table、K;
select n users from the user features information table and record them asU={u1, u2,
u3,. . ...,un} initialize K, flag M={m1,m2,. . .,mk}
Select highly active user feature vectors as K initial clustersheart, flag N={n1,n2,. . .nk}
Repeat For all ui∈U
For all nj∈N
nj∈max sim)ui,nj(
For all mi∈M
For all uj∈U
Until K no change
back

Therefore, the nearest neighbor is searched in the cluster to generate recommendation for the target user, which reduces the complexity of recommendation time.

3 FUSION USER ATTRIBUTE FEATURE SIMILARITY ALGORITHM

3.1 Traditional similarity calculation method

1) Pearson correlation coefficient, the value range is between [-1, 1], the larger the value, the greater the similarity, and the negative correlation is less meaningful for recommendation. The similarity between user a and user b is as follows:

$$sim^*(a,b) = \frac{\sum_{p\in P}\left(r_{a,p}-\bar{r}_a\right)\left(r_{b,p}-\bar{r}_b\right)}{\sqrt{\sum_{p\in P}\left(r_{a,p}-\bar{r}_a\right)^2}\sqrt{\sum_{p\in P}\left(r_{b,p}-\bar{r}_b\right)^2}} \quad (1)$$

$\bar{r}_a$ and $\bar{r}_b$ represent the average value of user a and user b's ratings for all items; $r_{a,p}$ and $r_{b,p}$ represent user a's rating of item p, and user b's rating of item p, respectively.

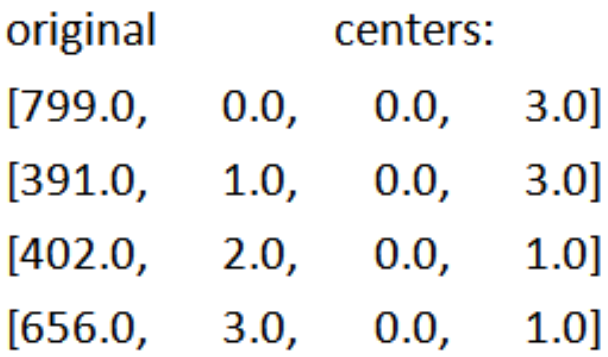

Figure 1: K value based on user activity

3.2 Time factor

The traditional collaborative filtering algorithm only considers the user's historical score record of the item when calculating the similarity between users, and ignores the dynamic change of user's interest in the item ,in the current various user behavior data sets, the time of user behavior is usually accurately recorded by the system, so it can be considered that the user's interest in a certain item will be affected by the time factor.Generally speaking, the items recently visited by users can more accurately reflect the current interest featuresof users.Therefore,this paper introduces the time factor to enhance the collaborative recommendation ability of users' recently visited items, and the relevant definitions are as follows.

Definition 1 tfirst:the time when the user first rated the item

Definition 2 ti: the time when the user rated product i

Definition 3 tall:the total time the user has used the system

According to the analysis, the user's interest in a certain item will gradually weaken over time and tend to a stable value,that is to say, there is a nonlinear negative correlation between user interest and time, which is similar to the forgetting law found by psychologist Ebbinghaus [12],Therefore, this paper uses the forgetting rule to quantify the time factor, and the formula for calculating the time factor weight of item P by user a is as follows:

$$\chi_p^a = e^{\frac{t_{first}-t_i}{t_{all}}} \tag{2}$$

Then combine the time weight with the Pearson similarity calculation formula,the formula is as follows:

$$sim^*(a,b) = \frac{\sum_{p\in P}\left(r_{a,p}\times\chi_p^a-\bar{r}_a^x\right)\left(r_{b,p}\times\chi_p^b-\bar{r}_b^x\right)}{\sqrt{\sum_{p\in P}\left(r_{a,p}\times\chi_p^a-\bar{r}_a^x\right)^2}\sqrt{\sum_{p\in P}\left(r_{b,p}\times\chi_p^b-\bar{r}_b^x\right)^2}} \tag{3}$$

Among them, $\bar{r}_a^x$ and $\bar{r}_b^x$represent the average score of any item of users a and B multiplied by the time weight.

3.3 Fusion user feature similarity algorithm

When calculating the similarity between users,the traditional similarity calculation method usually ignores the similarity of different user attributes.In user properties,age and occupation are the most influential attributes of project demand,therefore, this paper integrates the user's age and occupation features into the traditional similarity calculation method.

1) Age attribute similarity.The age similarity between user a and user B is as follows:

$$G(a,b) = e^{-|ga-gb|} \tag{4}$$

In the formula: G(a,b) ranges between [0,1], the greater the value, the greater the similarity;ga is the age of user a; gb is the age of user b.

2) Similarity of occupational attributes.The similarity of occupational attributes between user a and user B is as follows:

$$O(a,b) = \begin{cases} 1, o_a = o_b \\ 0, o_a \neq o_b \end{cases} \tag{5}$$

Where Oa is the occupation of user a; Ob is the occupation of user B.

3) User attribute similarity.Combining the above age attribute similarity and occupation attribute similarity,the similarity of user attributes is as follows:

$$sim^{GO}(a,b) = \alpha G(a,b) + (1-\alpha)O(a,b) \tag{6}$$

Where $\alpha \in [0,1]$ is the weight coefficient of the attribute,in different recommender systems, it can be adjusted by regression analysis fitting.

Combining the user attribute similarity with the Pearson correlation coefficient of the fusion time weight (formula 3), a new similarity calculation model can be obtained, that is, the fusion user feature similarity meter model, as follows:

$$sim^{**}(a,b) = \beta sim^{GO}(a,b) + (1-\beta)sim^*(a,b) \tag{7}$$

where $\beta \in [1,0]$ is the weight coefficient.Experimental results show that the model effectively solves the cold start problem of new users.

Table 2: Comparison of MAE values of different algorithms

Nearest neighbors KN	Algorithm				
	this paper	KUC	CoC	CFDP-CF	CFUI-CF
10	0.795	0.802	**0.792**	0.844	0.793
15	0.784	0.818	0.791	0.817	**0.781**
20	**0.761**	0.813	0.781	0.776	0.776
25	**0.732**	0.807	0.776	0.747	0.742
30	**0.731**	0.802	0.784	0.741	0.747
35	0.741	0.803	0.788	**0.740**	0.746
Average	**0.757**	0.808	0.785	0.778	0.764
WinningNum	3	0	1	1	1

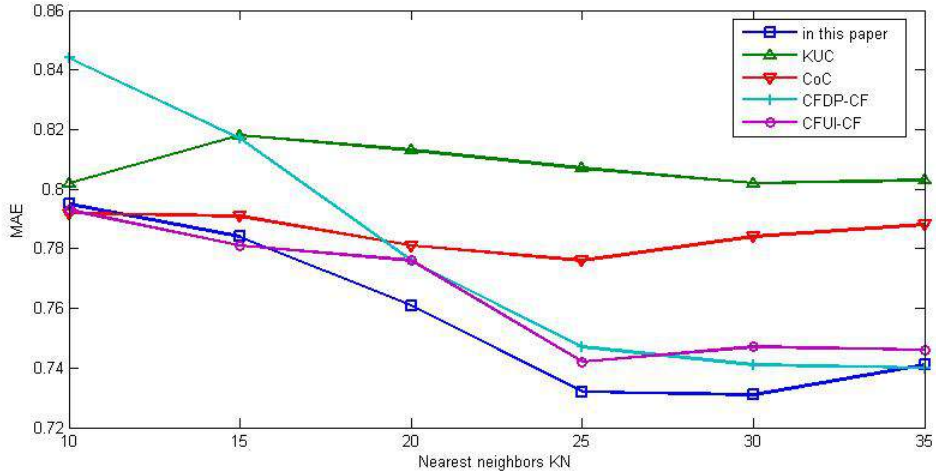

Figure 2: Comparison of mean absolute error

4 EXPERIMENT AND ANALYSIS

4.1 Experimental data

This experiment uses the MovieLens data set provided by the GroupLens group, which mainly includes three types of data sets with sizes of 100K, 1M, and 10M. The data set is open and used in experiments by many scholars.This paper mainly uses 1m data, which includes more than 6000 users' ratings of more than 4000 movies, as the experimental data set.

4.2 Algorithm comparison

In order to verify the performance of the proposed algorithm, the proposed algorithm will be tested on the data set movielens-1m,during the test, Mae and RMSE values will be compared with the following recommended algorithms:

1) K-meansUserClu)KUC(.One-sided user clustering algorithm uses K-means clustering algorithm to cluster users.

2) In [13], a two-sided user clustering algorithm is proposed, which combines users and projects, namely CoClust (COC).

3) The collaborative recommendation algorithm based on CFDP and time factor proposed in [14], namely CFDP-CF.

4) The collaborative recommendation algorithm based on user clustering and Slope One filling proposed in [15], namely CFUI-CF.

The comparison of MAE values of different algorithms is shown in Table 2.

In Table 2, the average value of MAE corresponding to different nearest neighbor numbers KN is represented by Average,and different nearest neighbors KN corresponds to the number of wins of different algorithms is represented by WinningNum.It can be seen from table 2that the performance of the algorithm proposed in this paper is better than that of other algorithms, whether measured by Average or WinningNum.The corresponding comparison chart is shown in Figure 2

As can be seen from Figure 2, the average absolute error of the algorithm proposed in this paper is lower than KUC algorithm and CoC algorithm on the whole, indicating that the recommendation performance of the algorithm proposed in this paper is better than KUC algorithm and CoC algorithm on the whole.When the value of KN is small, the recommendation quality of this algorithm is comparable to that of CFDP-CF algorithm and CFUI-CF algorithm, but with the increase of KN value, the performance of the algorithm in this paper is higher than that of CFDP-CF algorithm and CFUI-CF

algorithm,which indicates that the proposed algorithm has better recommendation performance in the case of large data sets.

5 CONCLUSION

This paper proposes a collaborative filtering recommendation algorithm based on hybrid clustering and fusion of attribute features, which combines user attribute feature with traditional similarity calculation method while clustering users.The experimental results have proved that both in the comparison of the MAE and the RMSE have obtained certain advantages,to a certain extent, it can improve the recommendation efficiency and solve the cold start problem, which opens up a new idea for the research of clustering algorithm in recommendation system.

REFERENCES

[1] WANG Guoxia, LIU Heping. Survey of personalized recommendation system[J]. Computer engineering and applications, 2012, 48(7): 66-76.
[2] BOKDE D K, GIRASE S, MUKHOPADHYAY D. An item-based collaborative filtering using dimensionality reduction techniques on mahout framework[J]. arXiv: 1503.06562, 2015.
[3] LEE D D, SEUNG H S. Learning the parts of objects by non-negative matrix factorization[J]. Nature, 1999, 401(6755): 788-791.
[4] Zhao Wei, Lin Nan, Han Ying, *et al.* An improved k-means clustering collaborative filtering algorithm [J]. Journal of Anhui University (NATURAL SCIENCE EDITION), 2016, 40 (2): 32-36
[5] Li Yanjuan, Niu Mengting, Li linhui. Collaborative filtering recommendation algorithm based on bee colony K-means clustering model [J]. Computer engineering and science, 2019, 41 (6): 1101-1109
[6] SUN Tianhao, LI Anneng, LI Ming, et al. Study on distributed improved clustering collaborative filtering algorithm based on Hadoop[J]. Computer engineering and applications, 2015, 51(15): 124-128.
[7] HU Xun, MENG Xiangwu, ZHANG Yujie, et al. Recommendation algorithm combing item features and trust relationship of mobile users[J]. Journal of software, 2014, 25(8): 1817-1830.
[8] LIU Haifeng, HU Zheng, MIAN A, et al. A new user similarity model to improve the accuracy of collaborative filtering[J]. Knowledge-based systems, 2014, 56: 156-166.
[9] WEI Suyun, XIAO Jingjing, YE Ning. Collaborative filtering algorithm based on co-clustering smoothing[J]. Journal of computer research and development, 2013, 50(S2): 163-169.
[10] ELKAHKY A M, SONG Yang, HE Xiaodong. A multi-view deep learning approach for cross domain user modeling in recommendation systems[C]//Proceedings of the 24th International Conference on World Wide Web. Florence, Italy, 2015: 278-288.
[11] CHEN Kehan, HAN Panpan, WU Jian. User clustering based social network recommendation[J]. Chinese journal of computers, 2013, 36(2): 349-359.
[12] Ebbinghaus forgetting curve [J]. Chinese nursing management, 2019, 19 (6): 860
[13] UEORGE T, MERUGU S. A scalable collaborative filtering framework based on co-clustering [C]// Proceeding of the 5th IEEE International Conference on Data Mining. Los Alamitos:IEEE, 2005:625-628.
[14] Zhang Kaihui, Zhou Zhiping, Zhao Weidong. Collaborative filtering recommendation algorithm based on CFDP and time factor [J]. Computer engineering and application, 2020, 56 (15): 80-85
[15] Gong min, Deng Zhenrong, Huang Wenming. Collaborative recommendation algorithm based on user clustering and slope one filling [J]. Computer engineering and applications, 2018, 54 (22): 139-143