Recommendation algorithm based on user attributes and tag preferences

Min Zhang\textsuperscript{1, a}, Xiaoling Yan\textsuperscript{1, b}, Hongwei Peng\textsuperscript{1, c}
\textsuperscript{1}Information Science and Engineering, Dalian University, Dalian 116000, China)
\textsuperscript{a}E-mail: zhangmin@dlu.edu.cn
\textsuperscript{b}E-mail: 1120834586@qq.com, \textsuperscript{c}E-mail: yxl19818965062@163.com, ‘E-mail: 2306118105@qq.com

Abstract—Aiming at the problem of sparse data and low recommendation accuracy of recommendation systems in a big data environment, a UT-CF algorithm that combines user attributes and tag preferences is proposed. The algorithm extracts and selects user attribute information, calculates the user attribute similarity matrix according to the user attribute characteristics; obtains the user’s tag score according to each score of the user item score matrix, obtains the user tag preference matrix, and calculates the difference between user tags Similarity, to get the user's similar neighbors; In order to reduce the search space of similar users, the K-means clustering algorithm based on the principle of maximum distance is introduced. Make the top k recommendations based on the recommendations of the target user's nearest neighbors. Through the experiment of MovieLens 100K data set, the algorithm improves the accuracy of recommendation.

1. INTRODUCTION
With the emergence and development of more and more large-scale commercial websites and shopping websites, users are also faced with more choices. However, in the face of a dazzling array of products, it is difficult for users to make a choice at a time. In order to reduce the burden of users’ choices, the recommendation system has become a popular and effective technology to solve information overload by virtue of its ability to quickly and accurately make choices that best suit users’ interests [1].

The idea of the collaborative filtering recommendation algorithm is to find similar neighbors with the same preference as the target user by analyzing historical score information, and make recommendations for the target user based on the historical score information and interests of the approximate neighbors [2]. However, in reality, it is not possible for every user to evaluate all items. User ratings for items are scattered sporadically, and the algorithm makes recommendations to target users based on the user's ratings for items, so the current algorithm faced with the problem of low recommendation accuracy due to too sparse data [3]. Regarding this problem, the literature [4] weights the confidence level and considers the user's personal interest in the similarity calculation formula, which improves the accuracy of the similarity calculation, but the recommendation result is relatively limited. The algorithm in [5] analyzes the item attributes and introduces a scale factor at the same time. It studies the relationship between the number of common and non-common ratings for the same item and the relationship between the item attributes, which improves the recommendation accuracy. Literature [6] combines technology and item clustering, and uses technology in different
categories to predict target users’ ratings by clustering items, but fails to grasp the important factor that users’ own characteristics also affect ratings. Literature [7] improved the matrix factorization scoring model, combining the matrix factorization based on item implicit feedback and the matrix factorization based on user implicit feedback with a certain weight to improve the accuracy of user prediction score, but the algorithm is more complicated. In [8], the author proposed a neural collaborative filtering method combined with deep learning. Although the recommendation effect is improved, the cost is high, which is not practical in real scenarios.

In response to the problems in the above-mentioned literature, this paper proposes a UT-CF algorithm that combines user tag rating preferences and user attributes. The algorithm uses user attributes and user tag rating preferences to calculate the nearest neighbors of the target user. At the same time, it uses the K-means based on the maximum distance principle for clustering, which can more accurately predict the preferences of the target user while reducing the dimension of calculation and improving the recommendation quality.

2. MATERIAL AND METHODS

2.1 User-based collaborative filtering recommendation algorithm

Recommendation algorithms based on collaborative filtering are mainly divided into user-based and item-based collaborative filtering recommendation algorithms. Among them, the user-based collaborative filtering recommendation algorithm is based on the same core idea that users with the same interest evaluate the item as the same [9].

The collaborative filtering recommendation algorithm process is as follows:

- Obtain user project rating data.
- Suppose there are m users and n projects.
- \( U = \{ u_1, u_2, ..., u_m \} \) and \( I = \{ i_1, i_2, ..., i_n \} \) to represent user collection items and \( r_{i,j} \) represent ratings.

| Table 1 Diagram of scoring matrix decomposition |
|-----------------------------------------------|
| \( u_1 \) | \( r_{1,j} \) | \( r_{1,2} \) | ... | ... | \( r_{1,n} \) |
| \( u_2 \) | \( r_{2,j} \) | \( r_{2,2} \) | ... | ... | \( r_{2,n} \) |
| . | . | . | . | . | . |
| \( u_i \) | \( r_{i,j} \) | \( r_{i,2} \) | ... | ... | \( r_{i,n} \) |
| . | . | . | . | . | . |
| \( u_m \) | \( r_{m,j} \) | \( r_{m,2} \) | ... | ... | \( r_{m,n} \) |

- Similarity calculation.
  - The traditional collaborative filtering recommendation algorithm judges whether the two have the same preference based on the similarity value. The higher the similarity value, the more similar the preferences of the two are, and the easier it is to become similar neighbors of the target user [10]. Whether the selection of similar neighbors is appropriate or not determines the accuracy of the recommendation with a high probability.
  - Find the similar neighbors of the target user and analyze the preferences of similar neighbors.
  - Predict the score and make recommendations.

Analyze the historical score information of similar neighbors, predict the target user’s score on other unrated items according to Equation (1), and recommend items for the target user according to the predicted score.

\[
r_{u,i} = \overline{r_u} + \frac{\sum_{v \in N_u} \text{sim}(u, v) (r_{v,i} - \overline{r_v})}{\sum_{v \in N_u} \left| \text{sim}(u, v) \right|}
\]  

(1)
2.2 K-means algorithm based on the maximum distance principle
Clustering algorithm is widely used, the core of which is to find a clustering center point, and to ensure that the square distance between each data and its nearest center point is the smallest [12]. Because the initial center point of the algorithm is prone to selection bias, the data originally in the same class is randomly selected as the initial center point, which makes the subsequent overall clustering data bias. Literature [13] adds the principle of maximum distance to the algorithm to maximize the distance between the initial center points of the cluster and avoid the initial center points from appearing in the same cluster. To improve calculation efficiency, the algorithm has introduced.

2.3 Recommendation algorithm based on user attributes and tag preferences (UT-CF)
The algorithm proposed uses the user attribute similarity matrix and the user’s tag rating preference matrix are introduced to improve the accuracy of finding similar neighbors; The K-means algorithm based on the maximum distance principle is introduced to cluster the user tag preference matrix to reduce the calculation range of finding similar neighbors of the target user.

2.3.1 Establish and analyze user tag preference matrix to obtain similar user sets
The rating matrix shown in Table 1 is composed of a list of user rating data tables for each item, and the rating information of a single user is as Table 2:

| Table 2 User -item rating |
|--------------------------|
| $u_i$ | $i_1$ | $i_2$ | ... | $i_l$ | ... | $i_n$ |
| 4     | 1     | ...   | 3   | ...   | 5     |

Each item in item set $I = \{i_1, i_2, ..., i_n\}$ has its own attribute or label. For example, the tags of Muppet Treasure Island are Adventure, Comedy, Musical, and Thriller. When the user rates the movie, it can be considered that the user has implicitly rated the movie’s label, and the implicit scoring of movie tags can dig out the user’s interest preference better than a single movie scoring.

When tags are involved in multiple items rated by the user, the average score of the user for the tags is taken. In the calculation, in order to avoid the deviation of the score due to the difference of the user’s personality and the scoring mood, this paper removes the maximum error value when calculating the average tag score to obtain a more accurate tag score. The tag’s average score see as Equations (2).

$$r_{\lambda} = \frac{\sum_{i \in I} \lambda r_{i\lambda}}{N_{\lambda}}$$  \hspace{1cm} (2)

Note: $I_{\lambda}$ is all items that the user has rated and also contain the label $\lambda$; $N_{\lambda}$ is the number of items with this label in $I_{\lambda}$; The value of $\lambda$ is 1 or 0.

The label rating preference matrix is shown in Table 3:

| Table 3 User label preference matrix |
|-------------------------------------|
| $u_i$ | $label_{1}$ | $label_{2}$ | ... | $label_{l}$ | ... | $label_{n}$ |
| $3.7$ | $3.1$ | $0$ | ... | $3.7$ |
| $2$ | $3.4$ | $4.2$ | ... | $3$ |
| ... | ... | ... | ... | ... |
| $u_i$ | $2.5$ | $3.6$ | ... | $4.7$ |
| ... | ... | ... | ... | ... |
| $u_n$ | $4$ | $2.8$ | ... | $4.3$ |
In order to reduce the calculation dimension, the K-means algorithm based on the principle of maximum distance is introduced to cluster the user tag rating preference matrix, find the cluster where the target user is located, and calculate the similarity with other members in the cluster to obtain the similar neighbor set $N_u$ of the target user.

2.3.2 Extract and select user attributes to establish similarity calculation formula to obtain similar neighbors

On the basis of ensuring accuracy, in order to increase the search breadth of similar neighbors, provide surprise recommendations for target users and solve the problem of boring recommendations. Introduce the third latitude (user attributes) to calculate user similarity to mine more items and recommend them to target users.

The occupation, age, and gender similarity between users are shown in Equations (3-5):

$$
sin(u,v)_{occ} = \sin(O_x, O_y) = \begin{cases} 0, & (O_x \neq O_y) \\ 1, & (O_x = O_y) \end{cases}
$$

$$
sim(u,v)_{gen.} = \sin(S_x, S_y) = \begin{cases} 0, & (S_x \neq S_y) \\ 1, & (S_x = S_y) \end{cases}
$$

$$
sin(u,v)_{age} = \sin(A_x, A_y) = \begin{cases} 10, & |A_x - A_y| > 10 \\ 1, & |A_x - A_y| \leq 10 \end{cases}
$$

Fused the three formulas to obtain the calculation formula of user attribute similarity.

$$
sim(u,v) = \alpha \sin(u,v)_{occ} + \beta \sin(u,v)_{gen.} + (1 - \alpha - \beta) \sin(u,v)_{age}
$$

By analyzing the user attributes and calculating the similarity, the user similar neighbor set $N_u$ is obtained.

Finally, after merging similar neighbor sets, predict the target user's score for the remaining items according to Equations (1) and take top-k recommendation.

3. Results

3.1 Data set

The MovieLens 100k data set is used as the data set for this experiment. The data set contains at least 100,000 ratings of 1682 movies by 943 users.

3.2 Evaluation index

The experiment used the mean absolute error (MAE) to test the algorithm recommendation effect, which can measure the accuracy of the recommendation. The smaller the value of MAE, the better the algorithm performance.

$$
MAE = \frac{1}{|V|} \sum_{v \in V} |p_{uv} - r_{uv}|
$$

Precision(Equation 8) is the ratio of the recommended successful data to the total data.; Recall(Equation 9) is the predicted ratio of the items to be recommended to be successfully recommended [15]. $R(u)$ is the recommended item set in the training set, $T(u)$ is the actual scoring item set.

$$
\text{precision} = \frac{\sum_{v \in V} |R(u) \cap T(u)|}{|R(u)|}
$$
The experiment used F1 index to test the accuracy and recall rate of the algorithm proposed in this article. F1 is introduced to verify the feasibility of the proposed algorithm.

\[
F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]  

(9)

\[
F_1
\]  

(10)

3.3 Optimal cluster number experiment

Figure 1 shows that 6 is the optimal number of clusters in this experiment.

After determining the optimal number of clusters, compare the MAE values of UT-CF algorithm and the collaborative filtering recommendation algorithm based on using pearson to calculate user’s similarity (UBCF), the collaborative filtering recommendation algorithm based on non-negative matrix factorization (NMF-CF), the collaborative filtering recommendation algorithm based on K-means clustering (Kmeans-CF). (see as Figure 2).

3.4 Top-k recommended accuracy experiment

Compare the F1 values of the UT-CF algorithm and UBCF, NMF-CF and Kmeans-CF algorithm (see as Figure 3) when the length of the recommended list is 30, 40, 50, 60 and 70.
4. DISCUSSION
The UT-CF algorithm constructed in this research combines machine learning technology and user information for user recommendations, and its accuracy is higher than related existing algorithms.

4.1 Optimal number of clusters in UT-CF algorithm
It is found from Figure 1 that the value is the smallest when \( k = 6 \), so the number of clusters is 6 is the best number of clusters in this experiment.

4.2 The accuracy of the UT-CF algorithm
Choose the number of nearest neighbors of \( u \) from 5 to 45 to compare with other algorithms. It can be seen from Figure 2 that the UT-CF algorithm will tend to be stable as the number of nearest neighbors of the target user increases, and the MAE value of UT-CF is smaller than other three algorithms. Taking the number of nearest neighbors 40 as an example, the MAE value of the UT-CF algorithm is 1.3%, 4.7% and 1.6% lower than the MAE values of UBCF, NMF-CF and Kmeans-CF algorithm respectively.

4.3 The recommended accuracy of UT-CF algorithm
It can be seen from Figure 3 that the UT-CF algorithm has a larger F1 value than the other three algorithms. When the recommended list length is 50, the F1 value of UT-CF reaches the highest point, but if increase the length of the recommended list, it can be seen that the F1 values of the UT-CF algorithm and other algorithms are showing a downward trend. In fact, the recommended list length is not the bigger the better, but there is a critical value. Comparing the F1 value of each algorithm when the recommended length is 50, it can be seen that the UT-CF algorithm is 20%, 11% and 5% higher than UBCF, NMF-CF and Kmeans-CF algorithm respectively.

5. CONCLUSION
Based on the results of the above experiments and analysis and discussion, the following conclusions are drawn:
- By analyzing users' ratings of item tags, users can get more precise interests and hobbies, and improve the accuracy of the recommendation algorithm.
- The idea of clustering and then calculating the similarity to obtain the nearest neighbor set of the target user reduced the search space of similar users.

ACKNOWLEDGMENT
This work was financially supported by the General Project of the Natural Science Foundation of Liaoning Province (20180551247) and the Key Research and Development Program of Liaoning Province (201704014).

REFERENCES
[1] Han S, Chee S, Han J, et al. RecTree: An efficient collaborative filtering method[C]//Proceedings of the Third International Conference on Data Warehousing and Knowledge Discovery. Springer-Verlag, 2001.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
[2] He Ming, Yao Kaisheng, Yang Peng, Zhang Jiuling. Collaborative filtering personalized recommendation based on the similarity of tag information features[J]. Computer Science, 2018, 45(S1):415-422(in Chinese).K. Elissa, “Title of paper if known,” unpublished.
[3] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(6):734-749.
[4] Zhou X, Shu W, Lin F, & Wang B. Confidence-weighted bias model for online collaborative filtering. Applied Soft Computing. doi:10.1016/j.asoc.2017.07.005.
[5] Li Shuzhi, Li Zhijun, Deng Xiaohong. Collaborative filtering algorithm combining scoring scale factor and item attributes [J]. Application Research of Computers, 2020, 37(03): 680-683 (in Chinese).

[6] You Haipeng, Li Hui, Wang Yunmin, et al. An improved collaborative filtering recommendation algorithm combing item clustering and slope one scheme [C] // Lecture Notes in Engineering & Computer Science, vol2215. 2015: 313-316.

[7] Li Kunlun, Su Huading, Rong Jingyue. A personalized recommendation system integrating projects and user implicit feedback [J]. Small Microcomputer System, 2020, 41(03): 519-525 (in Chinese).

[8] He X., Liao L., Zhang H., et al. Neural collaborative Filtering [C] // Proceedings of the 26th International Conference on World Wide Web, 2017: 173-182 (in Chinese).

[9] Su X., Khoshgoftaar T.M. A survey of collaborative filtering techniques. Advances in Artificial Intelligence, 2009, 2009: 4

[10] Wang Yunchao, Liu Zhen. A collaborative filtering algorithm integrating user preferences for items and attributes [J]. Computer Science, 2018, 45(S2): 412-416 (in Chinese).

[11] Herlocker L., Konstan A., Borchers S A., et al. An algorithmic frame-work for performing collaborative filtering [C] // Proc of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1999: 230-237.

[12] Xu D., Tian Y. A comprehensive survey of clustering algorithms [J]. Annals Data Science, 2015, 2(2): 165-193.

[13] Dong Liyan, Wang Yu, Ren Yi, Li Yongli. Collaborative filtering algorithm based on matrix decomposition and clustering [J]. Journal of Jilin University (Science Edition), 2019, 57(01): 105-110 (in Chinese).

[14] Koren Y., Bell R., Volinsky C. Matrix factorization techniques for recommender Systems [J]. Computer, 2009, 42(8): 30-37.

[15] Wang Yonggui, Song Zhenzhen, Xiao Chenglóng. Collaborative filtering recommendation algorithm based on improved clustering and matrix factorization [J]. Journal of Computer Applications, 2018, 38(04): 1001-1006 (in Chinese).