DE-BIASING RATING PROPENSITY ALGORITHM IN GROUP RECOMMENDATION

Junjie Jia, Tianyue Shang and Si Chen

Department of Computer Science and Engineering, Northwest Normal University, Lanzhou, Gansu, China

ABSTRACT

In recent years, group recommendation systems have gradually attracted attention with the increasing phenomenon of people's group activities. Nonetheless, most research focuses on optimizing machine learning models to fit user behavior data better. However, user behavior data is observational rather than experimental. Due to the different psychological benchmarks of user ratings, the training data evaluated by the algorithm cannot fully represent the real preferences of the target group. A De-Biasing Rating Propensity Algorithm in group recommendation is proposed. The proposed algorithm identifies user groups with similar behavior preferences through the Predict & AHC algorithm based on cosine similarity, and calculates user bias information by group and user preference tendency for different user groups. The De-Biasing Proportion on different items is used to build a rating bias consistency model, which effectively adjusts the user's predicted rating. The experimental results show that the algorithm can significantly improve the quality and fairness of recommendation.

KEYWORDS

Preference propensity, Evaluation bias, Fairness, Group recommendation

1. INTRODUCTION

With the rapid development of Internet technology and the increase of people's dependence on the Internet, the Internet provides people with massive information resources. In this context, it is difficult for people to find the content they are really interested in, and it is easy to fall into the dilemma of information overload. Recommendation Systems (RS) , as a prominent application of network individualization to solve the problem of information overload, have been widely used in e-commerce, news portals, content sharing platforms, social media and many other fields [1][2].At the same time, in real life, users always tend to consume in various groups, such as dinner parties, watching movies, and group travel. This requires the recommendation system not only to consider the individual needs of a single user, but also to make the recommendation results meet the expectations of different users as much as possible, so the Group Recommendation System (GRS) came into being[3].

In recent years, a lot of research achievements have been made on group recommendation. Due to the difference of the recommended objects, the goal of RRS is to help multi-user groups quickly screen a large amount of interactive information, balance the differences in the preference behavior of each user, and recommend products or services of interest to the group, reducing differences unnecessary bias and conflict among users. In order to improve the accuracy of group recommendation results and users' satisfaction with group recommendation list, most studies mainly solve the following two problems: the preference fitting of a single user and the preference fusion of group members. We believe that the user's historical behavior data

David C. Wyld et al. (Eds): ICAITA, CDKP, CBW, SOFT, CMC, Signal, MLT-2023
pp. 99-116, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130608

accurately describes the user's intention and is the main basis for user preference fitting, but it ignores the inherent subjective color of user behavior data and the implicit deviation between user and item matching information. Studies have shown that user behavior data is observational rather than experimental [4], so there will be various data biases. Directly fitting the data with the model and ignoring the inherent biases will lead to poor performance, to a certain extent, also damage the users' experience and trust of there commended service[5], therefore, among the many key problems faced by the recommendation system, the problem of data bias has gradually become an important factor restricting the development of the recommendation system technology[6].

Data bias research is the basis for ensuring the fairness of group recommendation systems. On the one hand, for users, users do not evaluate all items, but tend to select and evaluate items that they are interested in. Therefore, the user's rating data is usually missing, and there is a bias in the user preference fitting process. The problem of generalization makes the potential interests of users undiscovered, and the user's sense of fairness is reduced. On the other hand, for items, users are more likely to evaluate those items with higher or lower public ratings, but do not give feedback on some interesting, unpopular or new items, so that new items cannot be acquired , which is unfair to businesses that produce or sell high-quality items [7].In order to ensure a fair interaction between users and items, eliminating the rating bias has become a new direction in the research field of RS.A possible solution is to approximate the average rating by removing bias in ratings given by users. Some researchers have tried to solve this problem by analyzing the sentiment level of review texts using various classification methods[8], however, these techniques involve high computational complexity, which makes the performance of the system lag.

In this paper, a De-biasing Rating Propensity algorithm (DBRPA) in group recommendation was proposed. Based on the cosine similarity between user rating vectors, an agglomeration hierarchical clustering algorithm was applied to the complete user item matrix containing user preferences and predicted ratings to identify user groups with similar behavioral preferences. The effectiveness of the prediction and clustering algorithm on different group recommendation methods and clustering techniques was verified. Then according to "individual preferences and group preferences are correlated" to quantify the deviation of group preference tendencies and user preference tendencies, the de-biasing proportion (DBP)value of specific user is given. These DBP values will adjust the predicted rating vector of the user, so that the user's rating benchmark is consistent with the members in the group, further improving the recommendation quality. Experimental results show that the recommendation quality of the proposed algorithm is better than that of the existing group recommendation algorithms on the benchmark datasets.
The main contributions of this paper are as follows:

Firstly, we point that the existing data imbalance problem of GRS, the group members form false preference characteristics due to the influence of factors in the preference construction environment makes the data fitting, reflecting the inconsistency of users' rating psychology, similarly, the missing values of the original rating matrix have data bias problem which leads to lower recommendation accuracy.

Moreover, a predictive rating correction model is proposed that discovers groups by prediction and agglomeration hierarchical clustering algorithm, based on the predictions of a particular user evaluation data and the user's rating characteristics, the DBP value of the particular user is given to fit the predicted rating of the user, so as to make the rating benchmark of members in the group consistent and further improve the quality of recommendation.

Finally, we present a general de-biasing frame work to mitigate rating propensity bias in GRS, which takes into account both user consistency and group preference bias to measurebias

information, and combines different rating criteria for each user in the group to improve recommendation accuracy and fairness.

The rest of the paper is organized as follows. Section 2 introduces the related work in this paper; Section 3 analyzes the problem of data bias in group recommendation systems; Section 4 focuses on describing the DBRPA algorithm, including the theoretical basis and implementation steps; Section 5illustrates the performed experiments; Section 6contains conclusions and future work.

2. RELATED WORK

Recommendation systems (RS) can affect users' purchasing behavior and bring economic benefits, and have achieved good effect in different application scenarios. With the great success of the research on personalized recommendation systems, group recommendation systems, as an extension of them, are receiving more and more attention from researchers. For systems with explicit feedback (i.e., numerical rating), collaborative filtering (CF) models have done a lot of work in the past two decades for their accuracy and scalability. CF generates recommendations based on the assumption that "users' feedback behaviour always strongly reflects their true preferences", leveraging users' historical rating behaviour to quickly predict potential user-interested content. However, this assumption does not always hold. Existing group recommendation methods usually assume that rating data is randomly missing and unbalanced.

Group recommendation system (GRS) need to obtain unknown preferences based on the user's history, and therefore require a large amount of historical data from the user. For new users and new items, there is no corresponding historical rating data, which results in that the proportion of users' ratings in the rating matrix is too small, and the similarity between users and items cannot be calculated, thus the ratings cannot be predicted and the recommendation results for new users and new items cannot be obtained. There are many recommendation studies on sparsity issues, among them, Ghazarian et al. in [9] used support vector machine regression to train models that predict the missing values of the rating matrix by calculating the similarity of item features. Rendle combined the advantages of SVM with the factorization model and proposed an algorithm called Factorization Machine (FM) [10]. This algorithm can solve the problem of huge sparsity that support vector machine can not usually solve, but has high time complexity for data with strong relational patterns. Zhou et al. in [11] recommended an incremental algorithm based on SVD, which recalculates the SVD of the original matrix to address sparsity issues and dynamic interests of users. Xiangshi Wang et al. in [12] combined the trust social network to correct the preferences of group members, but trust is usually difficult to obtain and therefore the method is not easy to implement.

In behavioral economics and decision-making research, the term bias is used in an agnostic manner to denote systematic patterns that deviate from normative or rational judgement criteria [13].A bias term is a situation where a group is treated less favorably by the algorithm. Therefore, fairness is highly relevant in recommendation systems, for example, there may be bias against certain user groups, item categories, etc. One of them, Kamishima et al. in [14] developed a regularization method for enhancing recommendation independence. The regularization term in the form of a probability matrix decomposition to penalize the discrimination of the classifier, aiming to eliminate bias in the model construction process. This concept can be extended to logistic regression classifiers and various other probabilistic models. These methods have been successfully applied to the fair recommendation of films that contain sensitive attributes such as ethnicity or gender. From the perspective of biased data input, Kamiran et al. [15] proposed to preprocess training data before learning classifiers and change the data to eliminate discrimination.

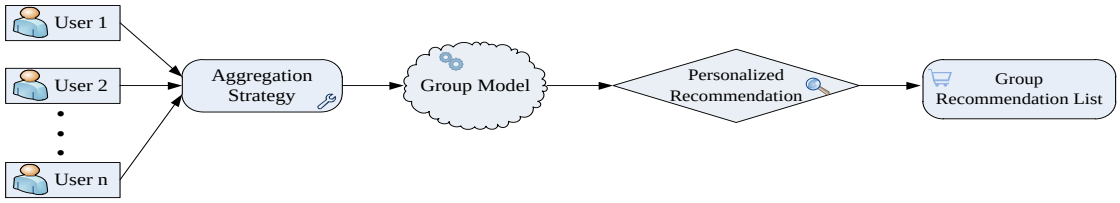
Researchers in behavioral decision-making, behavioral economics and applied psychology have found that people are often influenced by factors in the preference-constructed environment [16], such as current emotional state, interest in feedback processes, self-expression, etc., and have shown that missing values in the original rating matrix suffer from the same data bias problem. Calmon et al. in [17] proposed an optimization model to reduce discrimination probability by implementing data preprocessing through data probability transformation. The model defined discrimination and utility on a probability distribution, controlled data distortion on a sampling basis, and limited the impact of individual data transformation to ensure individual fairness. Thus, the discrimination control, data utility and individual data distortion are balanced in data preprocessing.

These methods overcome or reduce bias and discrimination in training data to some extent. But it can only preprocess the original training data, and has no generalization ability to process the unknown data. The method of overcoming or reducing bias by modifying training data only solves the problem from the perspective of statistical fairness, often only for a single sensitive attribute. Based on the above research inspiration, this paper considers from the prediction of the rating matrix to quantify the user bias information using the rating propensity, and attempts to remove the bias inherent in the rating matrix and the potential bias contained in the missing values, in order to establish a more accurate group recommendation model and achieve more efficient group recommendation.

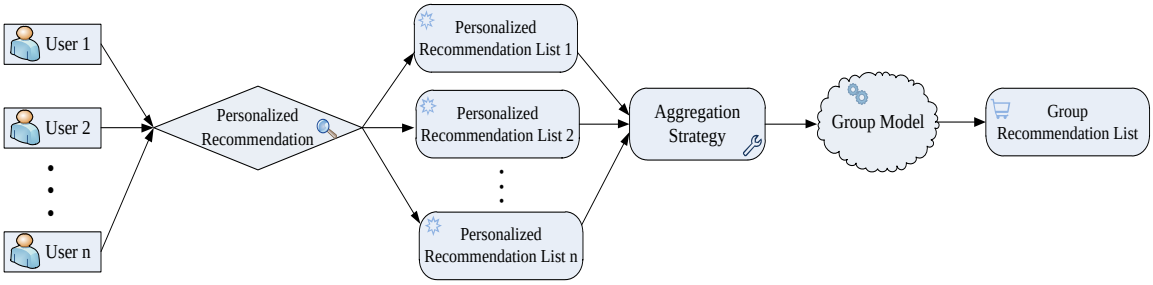
3. DATA BIASES IN GROUP RECOMMENDER SYSTEMS

Group recommendation systems (GRSs) are based on the assumption that "an item is uniformly liked by the users in a group, indicating that they all like the item". Based on each user's personalized preference vector, use preference aggregation strategies and preference aggregation methods are used to make each user agree with the group decision as much as possible.There are two main methods for fusing individual preferences into group preferences: model aggregation and recommendation aggregation [18] .The schematic diagram of the two fusion methods is shown in Figure 1.

- Model Aggregation: The behavioral preference characteristics (i.e. rating vectors) of each user in a group are fused according to anaggregation strategy to form a preference model for the group, which is then recommended to the group using a personalized recommendation system.
- Recommendation Aggregation: The personalized recommendation systems generate recommendations for each user and then fuse the recommendations of all users according to anaggregation strategy to form a group-oriented recommendation list.



a. Model Aggregation



b. Recommendation Aggregation

Figure 1.Diagram of the two aggregation methods

These two fusion methods need to fuse member preferences to construct group preferences, and the group preference characteristics formed by fusion can reflect the common interests of members in the group. Therefore, they are highly dependent on the user's historical rating behavior. The inherent bias in the user preference and the potential bias contained in the missing rating items have an impact on the final recommendation results. The mainstream aggregation strategies currently available are the average strategy [19], the least misery strategy [20] and the maximum pleasure strategy [21].When explicit feedback (i.e. explicit rating values) is used to fit preferences in user-line-based analysis, there are differences in how different users rate the same item, e.g. under the assumption that users have the same preference for a movie, they both rate it in the range [1, 5] but have different ratings (e.g. 3 and 4 respectively), mainly due to their own personalities or the rating environment they are in, which determines the tolerance they show when measuring things. In the case of group recommendation, it may be that users' ratings tend to be ambiguous or too extreme, so it is easy to obey the preferred behavior of most users in the group. Therefore, the recommendation results are not in line with their preferences, which reduces the satisfaction of users. At the same time, due to the influence of herd mentality, users may also make evaluations against their will, forming false preference characteristics and causing deviations in data fitting. Experiments show that the average strategy is more sluggish against these contradictory factors, while other fusion strategies are more sensitive, exposing fairness issues, such as the least misery strategy and the maximum pleasure strategy. Therefore, designing a reasonable and general de-biasing scheme is key to improving recommendations when fusion strategies are more sensitive to data bias.

4. DBRPA

The method proposed in this paper makes group members' preferences lean towards a consistent rating benchmark by analyzing the degree of bias of group members towards group preferences. In real life, the rating of the items we see is not only related to the user's level of interest in the item, but is also influenced by the user's own characteristics. Therefore, DBRPA is proposed in this paper, so that the group members have the same rating benchmark and further improve the quality of the recommendation. The framework flow of the de-biasing algorithm is shown in Figure 2.

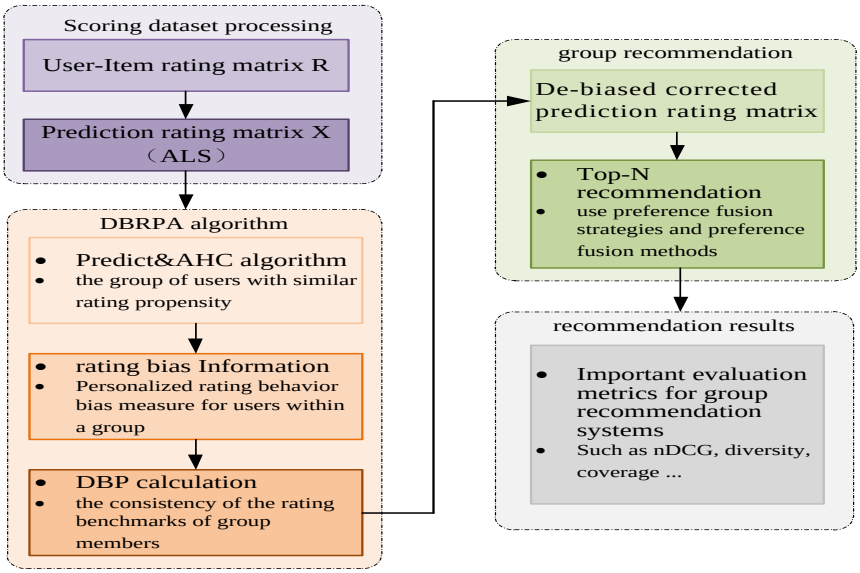


Figure 2. Process flow of DBRPA algorithm

4.1. Rating Prediction

Generally, the number of users in the rating datasets is far less than the number of items, so the sparsity of rating datasets is often very high. To reduce the impact of rating sparsity on the user clustering results, the rating matrix $\mathbf{R}^{m\times n}$ of m users and n items is decomposed into two low-dimensional user preference matrix $\mathbf{P}$ and item feature matrix $\mathbf{Q}$ using ALS matrix decomposition. Assign random initial values to $\mathbf{P}$ and $\mathbf{Q}$, and iterations are made so that the inner product prediction $\mathbf{X}$ of $\mathbf{P}$ and $\mathbf{Q}$ approximates the matrix $\mathbf{R}$. The user preference matrix represents the user's preference for individual item attribute characteristics, and the item feature matrix represents the affiliation of items to each attribute feature. $\mathbf{X}=\mathbf{P}\mathbf{Q}^T$, $\mathbf{P}\in\mathbf{C}^{m\times k}$, $\mathbf{Q}\in\mathbf{C}^{n\times k}$,and k indicates the number of features, generally k is much smaller than r ,and r denotes the rank of the matrix $\mathbf{R}$, $r\le\min(m,n)$.ALS has been proved to be effective in solving low-rank approximation problems and parallelization of large datasets. The objective function is as follows:

$$\min_{\mathbf{P},\mathbf{Q}} \mathrm{L}(\mathbf{P},\mathbf{Q})=\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\mathbf{R}_{ij}-\mathbf{P}_i\mathbf{Q}_j^T\right)^2+\lambda\left(\|\mathbf{P}\|_{\mathrm{F}}^2+\|\mathbf{Q}\|_{\mathrm{F}}^2\right) \quad (1)$$

where $\|\mathbf{P}\|_{\mathrm{F}}^2+\|\mathbf{Q}\|_{\mathrm{F}}^2$ is used as a regularization term to prevent over-fitting.

4.2. Group Discovery

Hierarchical clustering algorithms are considered as one of the most commonly used clustering algorithms because it does not need to set the number of clusters in advance and the ease of discovering hierarchical relationships between classes, while the calculation of cosine similarity is more reflected in the direction consistent similarity of user rating vectors. Therefore, the ALS matrix decomposition algorithm is used to predict the missing rates of individuals to obtain the complete user preference profile, and then a bottom-up hierarchical clustering method is used to partition the users into different categories. The similarity between two groups was defined as the

minimum similarity between users of two categories, and the algorithm terminated when all categories could not be combined. The cosine similarity is calculated as follows.

$$sim_{\cos}(u,v) = \frac{\sum_{i \in I} \left(\mathbf{X}_{u,i} \times \mathbf{X}_{v,i} \right)}{\sqrt{\sum_{i \in I} \mathbf{X}_{u,i}{}^2 \times \sum_{i \in I} \mathbf{X}_{v,i}{}^2}} \quad (2)$$

where $\mathbf{I}$ is the set of items; $x_{u,i}$ denotes the predicted rating value of user u foritem i and $x_{u,i}$ denotes the predicted rating value of user v for item i.

Suppose that the original rating matrix be R. The ALS matrix decomposition will be used to fill in the missing ratings to obtain the complete user rating vector and generate the user prediction rating matrix X, as follows.

$$\mathbf{R} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 5 & 4 & 3 & 2 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 \\ 4 & 3 & 2 & 1 & 5 & 5 & 5 & 5 \end{bmatrix} \quad (3)$$

Of these, u_4 was more forgiving and generally rated all the items he had experienced highly; u_3 was more critical and generally provided lower ratings.If the similarity is measured by the Pearson correlation coefficient, there is no correlation between u_4 and u_3. If the similarity is measured by cosine similarity and the similarity is 0.99, this paper concludes that although the rating behavior of users is inconsistent, the rating psychology is the same, that is, they express the same preference characteristics.

$$\mathbf{X} = \begin{bmatrix} 1.477 & 1.1630 & 0.9487 & 0.7260 & 1.0090 & 1.2317 & 1.4527 & 1.4361 \\ 4.9471 & 3.9873 & 2.9830 & 1.9789 & 1.0278 & 1.0118 & 0.9848 & 0.9609 \\ 0.9918 & 0.9938 & 0.9959 & 0.9851 & 0.9930 & 0.9918 & 1.0047 & 1.0092 \\ 3.9987 & 3.9995 & 3.9709 & 3.9674 & 3.9687 & 3.9723 & 3.9988 & 3.9911 \\ 4.0002 & 2.9731 & 2.0122 & 1.0472 & 4.9890 & 4.9705 & 4.9353 & 4.9951 \end{bmatrix} \quad (4)$$

The predicted rating $\mathbf{X}$ obtained by matrix decomposition is predicted based on a group of similar users, and this similarity is the similarity of preference characteristics and not the similarity of rating psychology. Therefore, this paper describes the difference of users' preferences from "dislike" to "love" according to the rating scale, and raises a question, that is, whether the same rating value has a consistent expression for different users' rating psychology? The rating bias problem that exists in the vacancy values of the original rating matrix. The user group obtained in this example is shown in Figure 3.

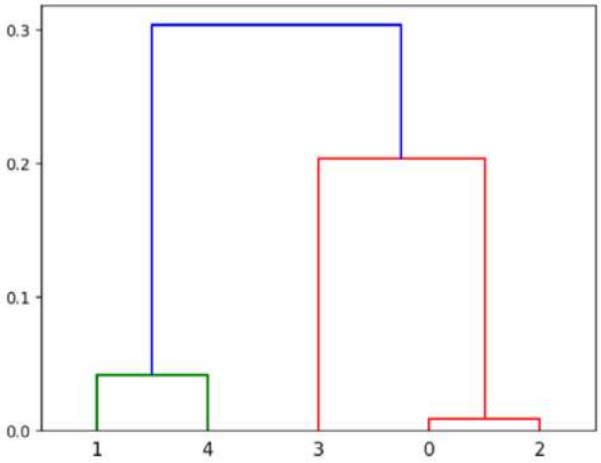


Figure 3. Agglomeration hierarchical clustering results

ALGORITHM 1: Predict&AHC Algorithm

Input：Original rating dataset

Output：Get the best group division result **G** according to user behavior preference characteristics

Step 1：Generate the original rating matrix **R** based on the user's original rating dataset information, where the user's unrated items are filled with 0.

Step 2：The ALS matrix factorization was used to fill in the missing ratings to obtain the complete user preference features, and obtain the prediction rating matrix **X** with the smallest objective function.

Step 3：Initialize the complete rating vector of each user as one cluster, and get n clusters in total.

Step 4：Determine the similarity between clusters based on SL , and obtain the initial similarity value of user preference behavior between clusters, namely single-linkage clustering method.

Step 5：Find the most similar user preferences between the two clusters and merge them into one cluster.

Step 6：Recalculate the preference behavior similarity between the new cluster and all clusters.

Step 7：Repeat steps 5 and 6 until the preference behavior similarity of all members in the cluster meets the minimum similarity threshold.

4.3. Modeling the Consistency of Rating Bias

In the context of RS, there is strong and consistent evidence from several studies that consumers provide display feedback that is biased towards system-generated recommendations [22]. However, the rating information provided by consumers is not solely related to the user's preferences, but also to the extent to which items are influenced by external ratings and the user's own characteristics, such as their natural optimism, current emotional state, and interest in the feedback process. Therefore, it is reasonable to think that consumers may be influenced by psychological factors to produce biased rating behavior. In this paper, the degree of rating bias of members in a group is defined by the user's personalized preference tendency and the group

preference tendency, that is, the degree of positive or negative relative to the group preference tendency. Finally, the de-biasing proportion value (DBP) for eliminating the bias of user ratings is determined based on the user rating vector to achieve unbiased

Definition 1 group preference propensity

Group preference propensity is a trend feature that aggregates the overall ratings of users on items within a group to indicate the expected value of ratings for a particular user group.

$$\bar{M}_{U_B} = \frac{1}{|U_B|} \sum_{u \in U_B} \frac{1}{|N|} \sum_{i \in I} \mathbf{x}_{u,i} \quad (5)$$

Let $\mathbf{I} = \{i_1, i_2, \cdots, i_n\}$ be a set of items, U_B is the group of users with similar rating trends, the number of users in the group is denoted by $|U_B|$ and the number of items is denoted by $|N|$. Where $\mathbf{x}_{u,i}$ is the user's predicted rating value for the item.

Definition 2 user preference propensity

The user preference propensityis the average of the overall predicted ratings of users without considering the item attribute characteristics to indicate the trend of personalized ratings of users.

$$\bar{M}_u = \frac{1}{|N|} \sum_{i \in I} \mathbf{x}_{u,i} \quad (6)$$

Definition 3 rating bias information

The user rating bias information is measured as the difference in response between the user's personalized rating behavior and the user group.If there is a group of similar users (similar rating trends/habits), the influence of psychological factors on users' rating behavior makesthe psychological benchmarks of user ratings are different, resulting in different preference characteristics among users. This bias phenomenon is called bias for user behavior.

$$bias_u^* = \bar{M}_{U_B} - \bar{M}_u \quad (7)$$

Where $\bar{M}_{U_B}$ indicates the rating feature of the group to which the user belongs for all items, and $\bar{M}_u$ indicates the rating feature of the user u.

4.4. De-Biasing Recommendations

Definition 4 DBP calculation

Let $DBP_u = \{DBP_1, DBP_2, \cdots, DBP_i, \ldots DBP_N\}$ be a set of DBP values,where DBP_i is the DBP value of this user for a single item i .For the target user, the DBP value reflects the weight proportion of the user to remove the bias for different rating items. As shown in Equation (8), the larger the value of the de-biasing proportion, the more the rating item deviates from the user's preference propensity.

$$DBP_i = \frac{\left(\mathbf{x}_{u,i} - \bar{M}_u\right)^2}{\sum_{u \in U_B} \left(\mathbf{x}_{u,i} - \bar{M}_u\right)^2} \quad (8)$$

The predicted rating of user u in the group is de-biased and corrected for item i according to the group preference tendency, as shown in Equation (9).

$$\hat{\mathbf{X}}_{u,i}^{*} = \mathbf{X}_{u,i} + DBP_i \square bias_u^{*} \square |N| \quad (9)$$

ALGORITHM 2: DBRPA Algorithm

Input：Prediction rating matrix $\mathbf{X}$

Output：De-biasing corrected prediction rating $\hat{\mathbf{x}}_{u,i}^{*}$

Step 1： $groups\ \mathbf{G} \leftarrow Predict\ \&\ AHC,\ U_B \subset \mathbf{G}$

Step 2：For U_B in $\mathbf{G}$:

Step 3： $\bar{M}U_B = \frac{1}{|UB|} \sum_{u \in U_B} \frac{1}{|N|} \sum_{i \in I} \mathbf{x}u,i$

Step 4：For u in U_B :

Step 5： $\bar{M}u = \frac{1}{|N|} \sum_{i \in I} \mathbf{x}u,i$

Step 6： $bias_u^{*} = \bar{M}_{U_B} - \bar{M}_u$

Step 7：For i in I :

Step 8： $DBP_i = \frac{(\mathbf{x}_{u,i} - \bar{M}_u)^2}{\sum_{u \in U_B} (\mathbf{x}_{u,i} - \bar{M}_u)^2}$

Step 9：End for；

Step 10： $\hat{\mathbf{x}}_{u,i}^{*} = \mathbf{x}_{u,i} + DBP_i \square bias_u^{*} \square |N|$

Step 11：End for；

Step 12：End for；

5. EXPERIMENTAL STUDIES

5.1. Datasets

To verify the feasibility of the proposed algorithm, the publicly available datasets MovieLens and Netflix are used for experimental verification. MovieLens is divided into several versions, and the specific information of ML-100K dataset used in this experiment is shown in Table 1. In this paper, all user rating data are divided into two parts in the experiment, with the training set accounting for 80% and the test set accounting for 20%. The algorithm model is trained in the training set and verified in the test set after the training is completed.

Table 1. Statistics of datasets

Datasets	#Users	#Items	#Ratings	Rating scale	Sparsity
ML-100K	943	1,682	100,000	[1, 5]	93.7%
Netflix	480,189	17,770	100,480,507	[1, 5]	98.8%

5.2. Evaluation Metrics

In this paper, the normalized Discounted Cumulative Gain (nDCG) is used as the evaluation index of recommendation accuracy. The same or similar items are always recommended to users in the actual recommendation, users will be resistant, which will lead to users' dissatisfaction with the recommendation results. Therefore, the diversity of recommendation list is equally important in group recommendation. The coverage of the recommendation results refers to the range of items covered by the recommendation results. High coverage means that the item types in the recommendation list account for a large proportion of all item types, which indicates that the recommendation system has a strong ability to explore potential item types and plays a positive effect on product sales. We use a satisfaction metric GSM to assess the fairness of the recommendation results by the group members. The fairness of the group is measured based on each member of the group. The formula is as follows:

$$\mathrm{GSM} = \frac{\sum I_u^s \cap I_r}{|N| \cdot |N_{ir}|} \quad (10)$$

Where $\mathbf{I}_u^s$ represents the items that members are satisfied with. $\mathbf{I}_r$ indicates recommended items. $|N_{ir}|$ indicates the number of recommended items.

5.3. Experimental results

In this paper, matrix factorization algorithm based on latent factor model (SVD) , Slope One algorithm for rating prediction model (SOP) [23] and DBT algorithm [24] are selected for experimental comparison. To test the performance of the proposed de-biasing algorithm for group recommendation, we conducted extensive experiments using various parameters, including the fusion strategy used, group size(P), and the recommendation list length(N).

5.3.1. Group Size

In order to study the recommendation performance of different algorithms under different group sizes and the influence of group discovery algorithms on the recommendation effect, the groups generated by Predict&AHC algorithm in this paper are compared with randomly selected groups of different sizes, and the recommendation list length is set to 10. The accuracy of group recommendation is compared when the group size is 5, 10, 15, 20, 25 and 30, and the average value is taken as the final result for five tests to ensure the fairness of the experimental results.

Table2.Experimental results of nDCG@10 with different group sizes in MovieLens

Algorithm	Randomly			K-means			Predict&AHC		
	5	10	20	5	10	20	5	10	20
SVD_AVG	0.153	0.176	0.159	0.801	0.807	0.791	0.897	0.878	0.859
SVD_LM	0.213	0.228	0.174	0.830	0.839	0.816	0.911	0.881	0.839
SVD_MP	0.164	0.157	0.102	0.811	0.785	0.738	0.883	0.823	0.771
SOP_AVG	0.162	0.215	0.158	0.812	0.816	0.769	0.887	0.877	0.853
SOP_LM	0.210	0.223	0.179	0.834	0.848	0.831	0.913	0.889	0.862
SOP_MP	0.175	0.159	0.127	0.826	0.796	0.765	0.887	0.843	0.773
DBT_AVG	0.173	0.237	0.202	0.828	0.830	0.806	0.9	0.891	0.869
DBT_LM	0.196	0.245	0.183	0.843	0.854	0.842	0.915	0.893	0.864

DBT_MP	0.192	0.168	0.125	0.837	0.816	0.768	0.897	0.855	0.778
DBRPA_AVG	0.182	0.213	0.158	0.851	0.876	0.850	0.892	0.883	0.869
DBRPA_LM	0.209	0.228	0.185	0.887	0.889	0.875	0.929	0.906	0.901
DBRPA_MP	0.197	0.216	0.176	0.878	0.877	0.853	0.893	0.897	0.878

1. normalized Discounted Cumulative Gain (nDCG)

In the experimental results shown in Table 2 and Table 3, Predict&AHC and DBRPA are the group recommendation algorithms proposed in this chapter to eliminate data bias.By analyzing the results in Table 3 and Table 4, it can be seen that: 1) The DBRPA algorithm performs better on the accuracy of unbiased recommendations for group members, which indicates the effectiveness of considering the psychological consistency and bias information of members in group recommendation. 2) The accuracy of recommendations obtained by using the SOP algorithm model for rating prediction combined with multiple aggregation strategies is significantly higher than that of the SVD model. This is because the SOP algorithm predicts the unknown preference based on the average deviation between items, and the addition of a new scoring term has a real-time impact on the prediction results.

Table 3.Experimental results of nDCG@10 with different group sizes in Netflix

Algorithm	Randomly			K-means			Predict&AHC		
	5	10	20	5	10	20	5	10	20
SVD_AVG	0.145	0.165	0.116	0.786	0.801	0.768	0.876	0.851	0.786
SVD_LM	0.224	0.214	0.186	0.821	0.798	0.786	0.893	0.889	0.861
SVD_MP	0.156	0.156	0.118	0.714	0.701	0.718	0.836	0.814	0.768
SOP_AVG	0.153	0.172	0.121	0.788	0.796	0.769	0.878	0.868	0.823
SOP_LM	0.206	0.217	0.178	0.826	0.802	0.788	0.901	0.889	0.870
SOP_MP	0.168	0.163	0.120	0.733	0.720	0.718	0.840	0.816	0.792
DBT_AVG	0.172	0.179	0.128	0.790	0.804	0.772	0.884	0.873	0.856
DBT_LM	0.218	0.239	0.179	0.837	0.816	0.790	0.903	0.885	0.873
DBT_MP	0.186	0.163	0.120	0.752	0.732	0.728	0.842	0.827	0.788
DBRPA_AVG	0.179	0.202	0.169	0.857	0.827	0.816	0.897	0.890	0.854
DBRPA_LM	0.214	0.205	0.163	0.874	0.826	0.804	0.915	0.903	0.885
DBRPA_MP	0.189	0.214	0.172	0.844	0.854	0.826	0.885	0.899	0.862

3) The group discovery algorithm proposed in this chapter shows high accuracy under the corresponding group size. The reason is that the algorithm proposed in this chapter overcomes the data bias problem existing in the missing data of scoring, and considers the scoring bias problem under the influence of adverse factors such as rating environment and user rating psychology. Mining the groups with more similar internal rating behavior, resulting in less conflicting preferences within clusters. This also indicates that although the recommendation accuracy is affected by undesirable factors such as user's rating environment and psychology, the effect is moderate compared to other recommendation algorithms.

2. Diversity & Coverage

For the MovieLens dataset, when the group size is greater than 20, the proposed algorithm in this paper is more diverse than other algorithms in terms of accuracy.With the continuous expansion of the group size,the algorithm in this paper makes the preference characteristics of each member treated fairly while reasonably eliminating data bias. As can be seen from Figure 4 and 5, compared with other comparison algorithms, DBRPA algorithm presents a steady increase in diversity and coverage with the increase of group size.However, DBT algorithm has poor performance in the experimental results. This is because the DBT algorithm identifies and removes cold users in the data pre-processing part, their preference characteristics are not considered, so the algorithm itself is biased.

From the experimental results in Figure 6 and Figure 7, we can see that the diversity and coverage of Netflix dataset in this algorithm are significantly higher than other algorithms. When the group size is 10 and 20, the coverage of the proposed algorithm is significantly improved compared with other comparison algorithms. Combined with Figure 4 to Figure 7, the proposed algorithm is able to perform better recommendation performance under different group sizes.

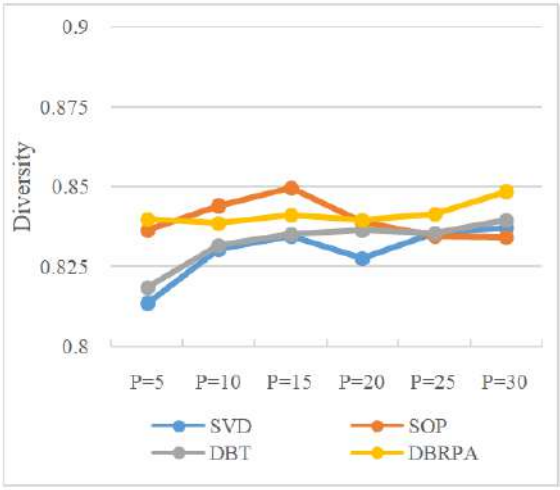


Figure 4.Comparison of diversity at different group sizes in MovieLens

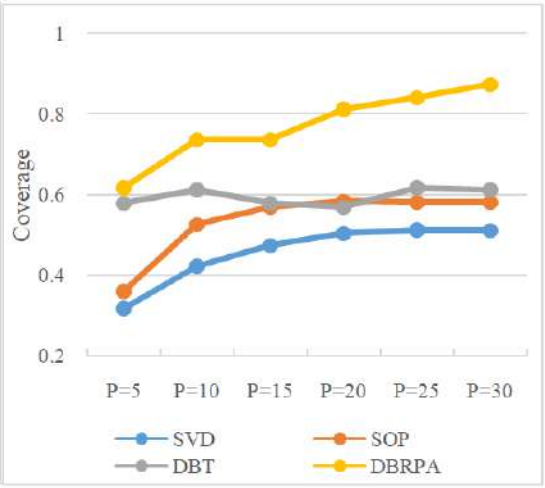


Figure 5:Comparisonof coverage at different group sizes in MovieLens

5.3.2. Recommendation List Length

In order to explore the influence of recommendation list length on group recommendation results, the recommendation results of each algorithm with different evaluation metrics are compared and analyzed.The fusion strategies used are: the average strategy (AVG), the least misery strategy (LM), and the maximum pleasure strategy (MP). The number of algorithm recommendations in the experiments is set to 5 respectively, and the rest of the parameters are the same as above for the following comparison experiments.

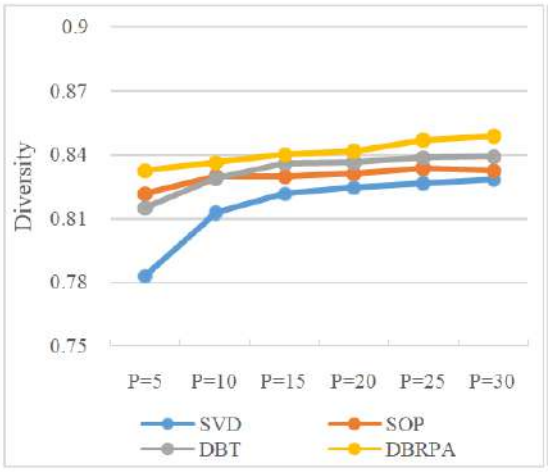


Figure 6.Comparisonof diversity atdifferent groupsizes in Netflix dataset

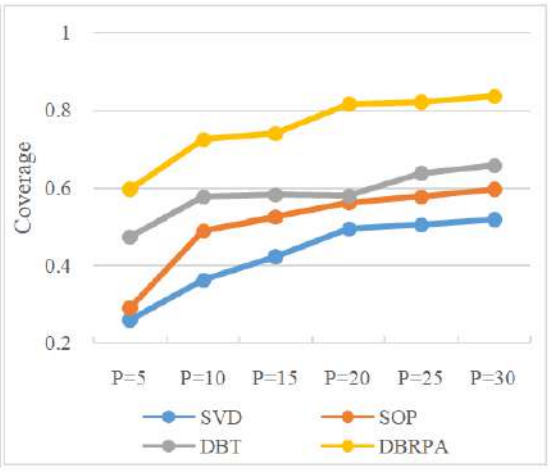


Figure 7.Comparisonof coverage atdifferent groupsizes in Netflix dataset

1. normalized Discounted Cumulative Gain (nDCG)

The recommendation results of each algorithm under different evaluation indexes in the MovieLens dataset are shown in Figure 8.It can be seen that the recommendation effect of the algorithm proposed in this paper is better than other methods under three preference fusion strategies, indicating that the introduction of group information with user rating similarity psychology can effectively improve the recommendation effect of group recommendation based on matrix decomposition. Comparing the recommendation accuracy (nDCG) results under AVG, LM and MP, it can be seen that the nDCG values of AVG and MP are generally higher than LM, which has better preference aggregation effect. This also indicates the influence of LM on recommendation accuracy due to biasing factors such as different psychological benchmarks for rating by members of the group.

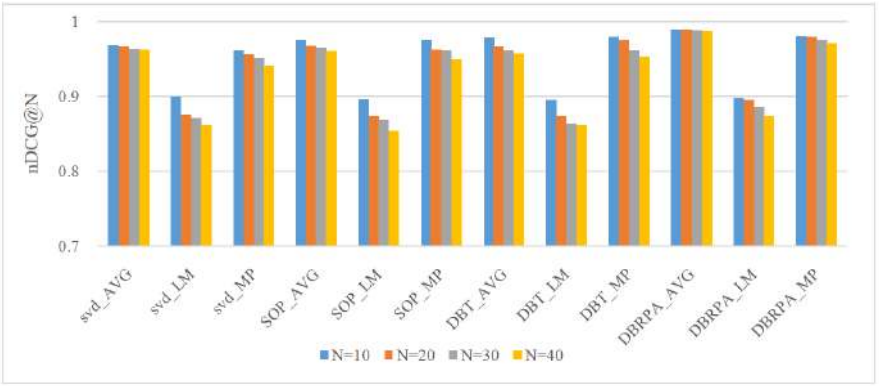


Figure 8. Comparison of recommendation accuracy in MovieLens dataset

2. Diversity & Coverage

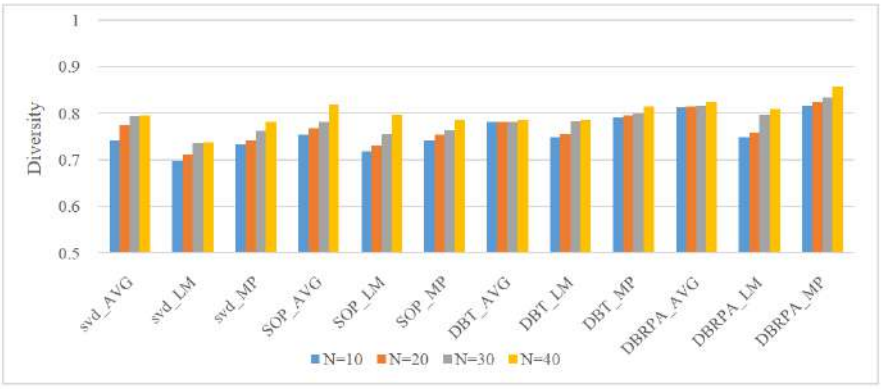


Figure 9. Comparison of recommendation diversity in Netflix dataset

The diversity and coverage of the recommendation list are usually at the cost of reduced accuracy, and the experimental results from Figure 9 show that when comparing different fusion methods, the level of diversity tends to have an opposite trend to the level of recommended accuracy. For the comparative experimental results of Netflix dataset, the algorithm proposed in this paper has higher diversity and coverage than SOP_LM and slightly higher than SOP_AVG and DBT when the number of recommendation lists is small, and its advantages become more and more obvious as the number of recommendation lists increases.The selection of predictive rating algorithm has the greatest impact on the recommendation coverage. The algorithm based on matrix decomposition always prioritizes the most popular and highest rated items, so the accuracy rate is high while the coverage rate is significantly lower than other algorithms.

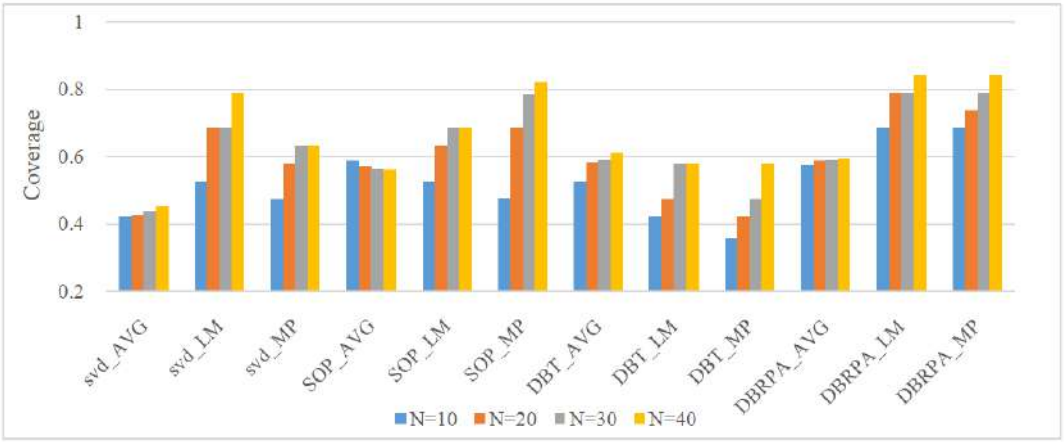
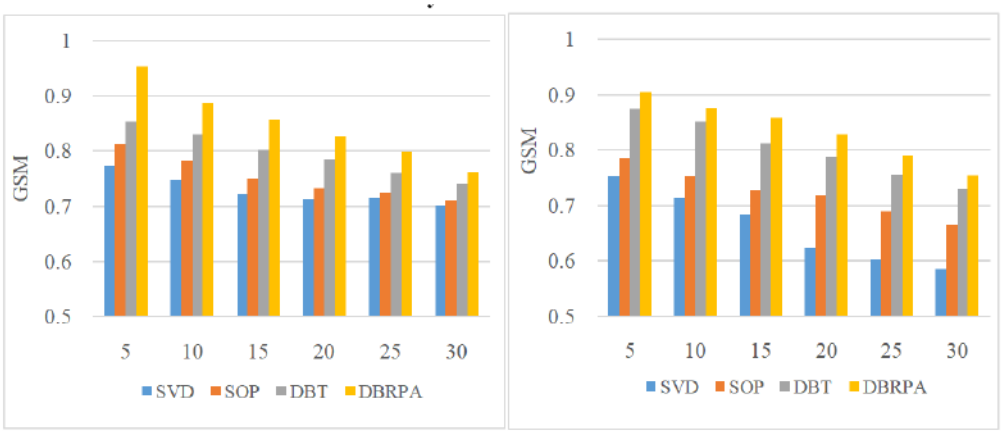


Figure 10. Comparison of recommendation coveragein MovieLens dataset

With the increase of the number of recommendations, the diversity and coverage of the recommendation list basically shows an increasing trend. Because when the number of recommendations is small, the preference needs of some members in the group are ignored and cannot be fairly recommended. The experimental results in Figure 10 show that the recommended coverage of the the proposed algorithm is still higher than the three benchmark algorithms under different aggregation strategies. The algorithm in this chapter not only considers the inherent bias of the rating data, but also deals with the consistency of the rating psychology of the members in the group. It enables each member's preference needs to be recommended fairly. The selection of predictive rating algorithm has the greatest impact on the recommendation coverage. The algorithm based on matrix decomposition always prioritizes the most popular and

highest rated items, so the accuracy rate is high while the coverage rate is significantly lower than other algorithms.

5.3.3. Recommendation Satisfaction Analysis

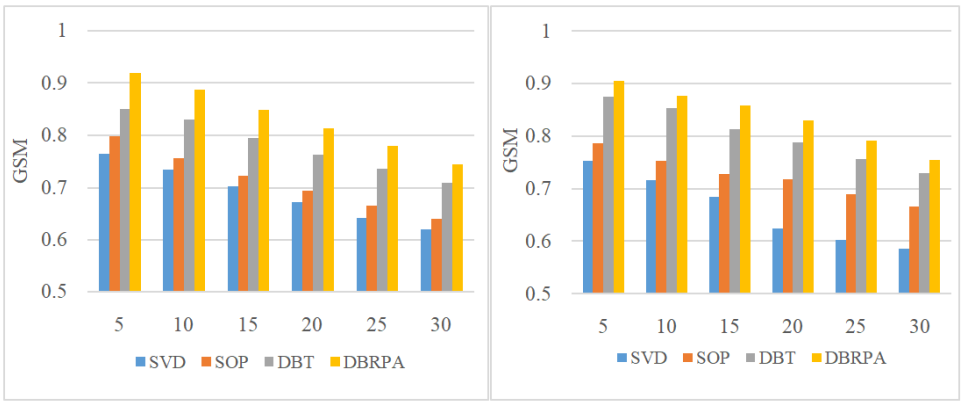


a. GSM contrastfigure for top-7 recommendation b. GSM contrastfigure top-11 recommendation

Figure 11. The GSM outcomes of methods in MovieLens dataste

GRSs aims to recommend suitable items to groups, meets the overall preference needs of group members as much as possible, and maximize the satisfaction of each member of the group. The validation of recommendation results is the main portion of the evaluation. The experiment adopts the average strategy to construct the group model.Compare DBRPA with SVD, SOP and DBT, and compare the member satisfaction of the items recommended by the algorithm in Top-7 and Top-11 respectively. The group rating is equal to the average rating of the members. The following comparative experiments are conducted in MovieLens and Netflix datasets.

The experimental results in Figure 11 show that when the group size of DBRPA is 5,10,15,20,25,30, the GSM values of Top-7 and Top-11 are higher than those of the other three algorithms. 1) The four algorithms have the highest GSM value when the group size is 5, and gradually decrease with the group size increases. This is because as the size of the group increases, conflicts between members of the group become more and more obvious. Preference aggregation is also becoming more difficult, which leads to lower satisfaction of group members with the recommendation results. 2) When the group size increases, the recommendation satisfaction decreases.



a. GSM contrastfigure for top-7 recommendation b. GSM contrastfigure for top-11 recommendation

Figure 12. The GSM outcomes of methods in Netflix dataset

As shown in Figure 12, when the group size of the algorithm DBRPA proposed in this chapter in the Netflix dataset is 5,10,15,20,25,30, the GSM values of Top-7 and Top-11 are higher than those of the other three algorithms. As the group size increases, the GSM value decreases but is higher than the other three baseline algorithms. In the Top-7 recommendation, when the group size of DBRPA is 5,10,15,20,25,30, the GSM values are 91.83%, 88.67%, 84.77%, 81.27%, 77.95% and 74.32%. In the Top-11 recommendation, when the DBRPA group size gradually increases, the GSM values are 89.36%, 86.27%, 83.86%, 81.35%, 78.76% and 75.2% respectively. In addition, the four algorithms have the highest satisfaction when the group size is 5. The reason is that the smaller the size, the easier it is for the recommendation results to meet the preference needs of members in the group.

6. CONCLUSIONS AND FUTURE WORK

This paper proposes the application of DBRPA in group recommendation. Theoretical research shows that the missing rating items in the matrix factorization model also have the problem of data bias, and DBRPA algorithm is used to alleviate the impact of this bias.Compared with the existing group recommendation algorithms, the experimental results show that these proposed methods have effective debiasing performance, and the proposed algorithm has better performance in multiple evaluation indicators such as accuracy.

In future work, we can analyze the recommendation generation process with the causal graph to improve the recommendation performance of the algorithm by considering information such as user ratings and group structure, and further analyzing the fusion of other social group information that is widely available. At the same time, user preferences may also change with the passage of time, we can study the de-biasing operation in dynamic scenarios.

REFERENCES

[1] Yanxiang Huang, Bin Cui, Jie Jiang, Kunqian Hong, Wenyu Zhang, and Yiran Xie. Real-time video recommendation exploration. In Proceedings of the 2016 International Conference om Management of Data (SIGMOD'16). ACM,2016：35-46.

[2] Aaron van den Oord, Sander Dieleman and Benjamin Schrauwen. Deep content-based music recommendation. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13), 2013, 2643-2651.

[3] Inmaculada Garcia, Sergio Pajares, Laura Sebastia and Eva Onaindia. Preference elicitation techniques for group recommender systems. J. Information Sciences, 2012, 189(11)：155-175.

[4] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2020. Bias and Debias in Recommender System: A Survey and Future Directions. arXiv Preprint (2020).

[5] Harald Steck. 2018. Calibrated Recommendations. In Proceedings of the 12th ACM Conference on Recommender Systems (RecSys'18). ACM, New York, NY, USA, 154-162.

[6] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. arXiv Preprint (2019). https://arxiv.org/abs/1908.09635

[7] Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'18). ACM, New York, NY, USA, 2018, 2219-2228.

[8] Chen Lin, Xinyi Liu, Guipeng Xu, and Hui Li. Mitigating Sentiment Bias for Recommender Systems. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21). ACM, New York, NY, USA, 2021, 31–40.

[9] Sarik Ghazarian and Mohammad Ali Nematbakhsh. Enhancing memory-based collaborative filtering for group recommender systems. J. Expert Systems with Applications, 2014, 42(7)：3801-3812.

[10] Steffen Rendle. Factorization Machines. In Proceeding of the 10th International Conference on Data Mining. IEEE, Sydney, NSW, Australia, 2010, 995-1000.

[11] Xun Zhou, Jing He, Guangyan Huang and Yanchun Zhang. SVD-based incremental approaches for recommender systems. J. Journal of Computer and System Sciences, 2014, 81(4)：717-733.

[12] Xiangshi Wang, Lei Su, Qihang Zhou and Liping Wu. Group Recommender Systems Based on Members' Preference for Trusted Social Networks. J. Security and Communication Networks, 2020, 1-11.

[13] Martie G. Haselton, Daniel Nettle and Paul W. Andrews. The Evolution of Cognitive Bias. J. The Handbook of Evolutionary Psychology, 2015, 724-746.

[14] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh and Jun Sakuma. Fairness-Aware Classifier with Prejudice Remover Regularizer, 2012, Vol. 7524. Springer-Verlag, New York, NY.

[15] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. J. Knowledge and Information Systems, 2012, 3(1)：1-33.

[16] Gretchen B. Chapman and Eric J. Johnson. Incorporating the Irrelevant: Anchors in Judgments of Belief and Value. J. Heuristics and biases: The psychology of intuitive judgment, 2002, 120-138.

[17] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy and Kush R. Varshney. Optimized pre-processing for discrimination prevention. In Proceeding of the 31st International Conference on Neural Information Processing Systems (NIPS 2017). Advances in neural information processing systems. Long Beach, CA, USA, 2017, 3995-4004.

[18] Sihem Amer-Yahia, Senjuti Basu Roy, Ashish Chawlat, Gautam Das and Cong Yu. Group recommendation: semantics and efficiency. J. Proceedings of the VLDB Endowment, 2009, 2(1)：754-765.

[19] Hamidreza Mahyar, Elahe Ghalebi K and S. Mojde Morshedi. Centrality-based Group Formation in Group Recommender Systems. In Proceeding of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee (IW3C2), ACM, 2017, 1187-1196.

[20] Akshita Agarwal, Manajit Chakraborty and C. Ravindranath Chowdary. Does order matter? Effect of order in group recommendation. J. Expert Systems with Applications, 2017, 82(3)：115-127.

[21] Ludovico Boratto, Salvatore Carta and Gianni Fenu. Discovery and representation of the preferences of automatically detected groups: Exploiting the link between group modeling and clustering. J. Future Generation Computer Systems, 2015, 64(10)：165-174.

[22] Gediminas Adomavicius, Jesse C. Bockstedt, Shawn P. Curley and Jingjing Zhang. Reducing recommender system biases: An investigation of rating display designs. J. MIS Quarterly: Management Information Systems, 2019, 43(4)：1321-1341.

[23] R. Barzegar Nozari, H. Koohi and E. Mahmodi. A novel trust computation method based on user ratings to improve the recommendation. J. International Journal of Engineering (IJE), 2020, 3(3)：377-386.

AUTHORS

JunJie Jia: male, 1974, Ph. D. , Associate professor, master supervisor, research direction is Data Mining and Privacy Protection.

Tian Yue Shang: female, 1998, Graduate student research direction is Recommendation System and Data Mining.

Si Chen: female, 1994, Graduate student research direction is Recommendation System and Data Mining.