

A Personalized User Recommendation Based On Attributes Clustering and Score Matrix

¹C.Balasubramanian, ²J.Raja Sekar, ³M.Shenbaga Devi

¹Assistant Professor, ²Associate Professor, ³PG Student, Department of Computer Science and Engineering
MepcoSchlenk Engineering College, Sivakasi, India

¹balasubramanian@mepcoeng.ac.in, ²jrsekar@mepcoeng.ac.in, ³devi27695@gmail.com

Abstract-The evolution of the internet has provided the available access to numerous items. Over the past few years, recommender systems have been more and more effective in helping people identify their preferred resources from a large number of candidate objects. RS assesses the preferences of users who have previously shown an interest in certain items, in order to recommend items which better match the target user's interests. Personalized RS helps people save a considerable amount of shopping time. Collaborative Filtering is one of the most frequently used RS techniques. With the increase of users and items, data sparseness and poor expansibility are the key factors that affect the quality of recommender. In order to solve these problems and find the neighbor users accurately, clustering technology is introduced to optimize collaborative filtering algorithm. Firstly, user attribute distance is calculated. Secondly, we fill the rating matrix based on data filling approach. Thirdly, we use the filled matrix to calculate the similarity of users who are in the same cluster. Next, the k -nearest neighbors are generated to predict scores. Finally, the Top- N interested items are recommended to the target user. we conduct a series of experiments in the Movie Lens Movie dataset. Experimental results show the effectiveness of our approach.

Keywords-Data Mining, Recommender System, Clustering, Collaborative Filtering, Rating Prediction.

I. INTRODUCTION

The rapid development of e-commerce has facilitated consumers to purchase products from online. Nowadays more products are introduced in the market. Peoples have more choices for buying a particular product which also consists of many types. Most e-commerce web sites like Amazon, flip kart, snap deal encourage patrons to rating for their purchasing

products. This gives large collections of online customer ratings on the Web. These comments have become helpful for both consumers and organization. As of late individuals have been getting increasingly digitized data from Internet, and the volume of data is bigger than whatever other point in time, achieving a state of data over-burden. To tackle this issue, the recommender framework has been made because of the need to scatter so much data. It doesn't just channel the commotion, additionally help to choose appealing and valuable data. Recommender framework has made beginning progress in view of a study that shows no less than 20 percent of offers on Amazon's site originated from the recommender framework. Clustering-based approaches offer an alternative to model-based methods. Instead of decomposing the rating matrix into matrices with small ranks, these approaches reduced the search space by clustering similar users or items together. For example, Face book, Twitter, Yelp, Douban, Epinions, and so forth. Recommender systems are emerging as tools of choice to select online information relevant to a given user. One simple way of quantifying a user's like or dislike for an item is to use a rating. This can take a binary form as in the case of "like" vs "dislike", or have more shades as found in a 1-to-5 Star rating scale. Regardless of how the metric is obtained, the underlying assumption is that a rating is a good enough representation of the true feeling of a user towards a particular item. And once enough ratings have been collected in a community of users, a Collaborative Filtering algorithm can start working to recommend target users other items that might be of interest to them. Many E-Commerce companies have already incorporated RS with their services.

Examples for such recommendation systems include Product and Book recommendation by Amazon, Movie recommendations by Netflix, Yahoo! Movies and Movie

Lens, Product advertisements shown by Google based on the search history. A vast number of ratings and opinions of users regarding to a particular product is available in internet nowadays and these reviews are very much useful for a person to whether go for a particular service or not, for example if a person wants to buy a particular product he/she will read the reviews and watch rating of the product and then he/she will come to know about the product, people will believe judgment of trusted users, since there will be a vast number of reviews it had become a tedious task to mine reviews and users relation in social networks and it is an important issue in natural language processing, and web mining. In these days E-Commerce industry is growing at the exponential rate. As the users are using these systems their changing needs and variety of products is making business more and more complex. At the same time, they are providing ease and flexibility to users. In such environment, customers have difficulty to find optimal information about products from the tremendous amount of information. To help the buyers, the major e-business companies are going to develop their Recommender System to help their customers to choose items more efficiently this serves win-win strategy in E-Commerce. Recommendation Systems are becoming tools of choice to select the online information relevant to a given user. The remainder of this paper is organized as follows.

In section II, we present the related work to our proposed system, In Section III, the proposed model and its functionalities is described thoroughly. Results obtained from the proposed work are given in Section IV. Finally section V concludes the proposed work.

II. RELATED WORK

Clustering algorithms are often used in the recommend systems to group the users (items) into different clusters. A two-dimensional matrix is often used as connection matrix to cluster users. Zhang et al [1] clustered users by k -means algorithm based on the statistics of user behaviors on different item genres, which improved the quality of the recommendation. The statistics of users behaviors on different item genres performed well in measuring the users' interests and clustering them. Sarwar et al [2] used the k -means method to cluster the users, and the cluster of the target user was taken as the nearest neighbor search space, which reduced the range of the nearest neighbor search and improved the expansibility problem.

C.L. Liao et al [3] proposed a clustering based approach which applied a self-constructing clustering algorithm to reduce the dimensionality related to the number of products. They considered both the centers and variances of clusters, resulting in a similarity measure better than that proposed in other methods. Due to dimensionality reduction on the number of products, the processing time for making recommendations was much reduced. C. Wang et al [4] provided the detailed

cluster formation and reputation distribution to illustrate the slight difference between the outcomes under three decision making criterions, in which the choice of learning objects was governed by their reputations. During the strategy evolution, the player's state was updated, that is, each player can only own a chance to update his strategy once on average within a full Monte Carlo simulation (MCS) step.

This new line of Recommender Algorithms is termed Model Based and makes use of matrix factorization techniques to guess at the values of latent factors, which loosely stated could be thought as underlying descriptors of the domain. Koohi et al [5] Fuzzy C-means clustering to user-based CF and showed that fuzzy clustering outperforms other clustering techniques. The idea of using a factoring strategy came up during the Netflix.com competition when one of the contenders described the idea in a blog post. Other participants quickly saw the potential, and put the proposal to work. At the end, the winning team confessed that it relied heavily on matrix factorization to achieve the 10% improvement goal set by the judges. While matrix factorization does currently have an upper hand on Recommender Systems, its benefits don't come for free. For starters, the algorithm is based on factoring an inherently sparse matrix, which can easily be proven not to have a unique solution. The factorization is a time consuming and computationally expensive operation, needing to be recomputed every time a new rating is entered. On the other hand, the complex model produced by this technique can be prepared online, and made ready for almost instantaneous recommendations to be made. This approach was first taken by Amazon.com which holds a patent to the Technology from 2003 [6].

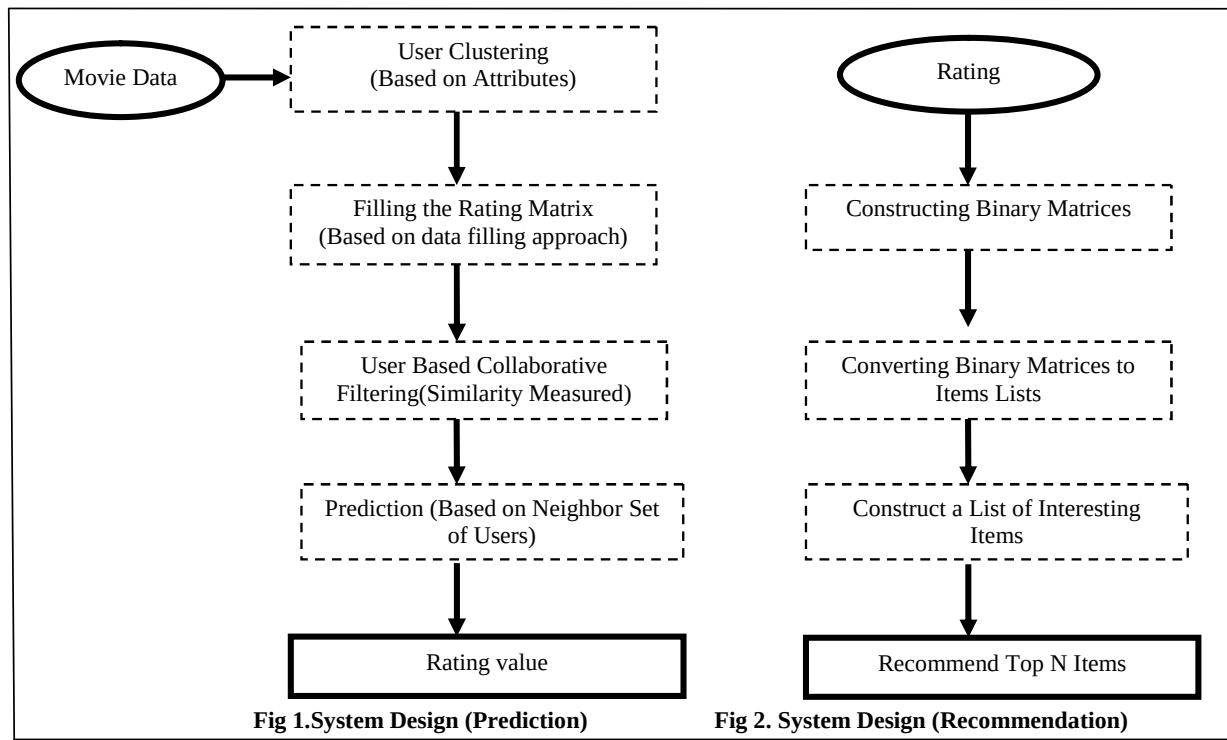
Most forms of Collaborative Filtering Recommender Systems work by matching a target user to a large list of candidates, by using a similarity function. This approach is termed User Based. The algorithm [7] takes the list of rated items from the target user, matches them to the rated items from a candidate user, and computes the squared error difference another suitable measure function, where the smaller the result, the closer the candidate's taste to the target user. The second broad category of Recommender Systems that has widespread use, is termed Content Based [8]. The group-based method was proposed to cluster users. Tian et al [9] presented a statistical model to estimate worker reliability and task clarity by characterizing the grouping behavior to form schools of thought.

Gao et al [10] proposed a group-based ranking method that evaluates users' reputations based on their grouping behaviors. Users were assigned with high reputation scores if they always fall into large rating groups. Their method was more accurate and robust than the correlation-based method in the presence of spamming attacks. Matrix filling technology has been widely used in many data mining tasks. In the recommender

system, the number of users and items is very large. However, every user just rates a few items, and every item is only rated by very few users. As a result, the rating data is extremely sparse leading to a miscalculation of the similarities of users and items, which then affects the prediction accuracy seriously. Filling the un-known ratings is the most direct way to density the sparse data thereby boost the performance of clustering.

It is crucial to select an appropriate filling algorithm. A dynamic evolutionary clustering model based on user attributes distance to cluster users. The similarity measure should be normalized and easily combined with other similarity measures. User-class score matrix, since the number of item class is far less than the number of items, which reduced the matrix dimension, increased the matrix data density and improved the cold start problem.

III. Proposed System



u_1 means user u 's characteristic value of age

$$u_1 = \begin{cases} 0, & \text{if the age of } u \leq 20 \\ 1, & \text{if } 21 \leq \text{the age of } u \leq 40 \\ 2, & \text{if the age of } u > 41 \end{cases}$$

u_2 means user u 's characteristic value of gender,

$$u_2 = \begin{cases} 0, & \text{if the user is male} \\ 1, & \text{if the user is female} \end{cases}$$

u_3 means user u 's characteristic value of occupation,

$$u_3 = \begin{cases} 0, & \text{if the user is in leisure class} \\ 1, & \text{if the user is in culture class} \\ 2, & \text{if the user is in management class} \end{cases}$$

The user distance d_{uv} is given as follows

$$d_{uv} = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + (u_3 - v_3)^2} \quad (1)$$

This section provides the detailed description of the proposed system. The system design is shown in Fig.1 and Fig.2. The dataset contains product rating which are collected from Movie lens website.

A. USER CLUSTERING BASED ON USER ATTRIBUTE DISTANCE

Usually, everyone has his own attribute, such as age, gender, native place, ID number, occupation, salary, and so on. Users of different age, gender, and occupation have different

Interests. According to statistics above, we divide age into three categories. When age is vectorized to three values, similar users can be better clustered, and the prediction results are better. In the dataset used in this paper from MovieLens, it includes 21 kinds of occupations. We separate occupations into 3 large categories depending on professionalism. They are leisure class, culture class and management class. In this paper, we analyze attribute characteristic of users to calculate user distance.

Here, $\{u_1, u_2, u_3\}$ are three attributes of user u ,

$\{v_1, v_2, v_3\}$ are three attributes of user v .

Clustering rule is generated based on distance of users. The smaller user distance d_{uv} is, the closer the user's interests are. Thus, similar users are divided into one community accurately.

B. FILLING THE RATING MATRIX BASED ON DATA FILLING APPROACH

In the recommender system, the number of users and items is very large. However, every user just rates a few items, and every item is only rated by very few users. The rating data is extremely sparse leading to a miscalculation of the similarities of users and items, which then affects the prediction accuracy seriously. Matrix filling technology has been widely used in many data mining tasks. The data filling was a simple and efficient way to improve the calculating effect of similarity and also the sparsity of the matrix is reduced. A RS consists of a set of M users U which will be indicated for short as the user-

set, a set of N items I named item-set. A collection of rating values expressing the preference of one user on a corresponding item. Such collection of preference indicators can be represented as a $M * N$ rating matrix. Where r_i^u is the rating given by the user u on the item i . Ratings can be integer values within a scale 1 (low interest) to 5 (strong interest). The rating for the pair (user, item) is unknown assume $r_i^u = 0$. $U(i)$ the set of users who evaluated the item i . $I(u)$ is the set of all the items for which the user u has expressed her preference. Filling approach to fill the matrix from both user and item sides, which overcomes the sparseness.

$$r_{ui} = \begin{cases} \text{if user } U \text{ rated item } I, \\ (2) r_{ui} \\ \text{else } r_{ui} \end{cases}$$

C. USER BASED COLLABORATIVE FILTERING

While User based Collaborative Filtering focuses on the relationship between rows. User-based CF firstly calculates the similarities between the active user and the other users. Cosine based Similarity Function to calculate the similarity between the users. Users with high similarities are selected as neighbors of the active user. It utilizes the neighbors' ratings on a specific item to calculate the weighted average which is used for predict the rating for the active user. The taste of items remains constant or change very slightly similar users build neighborhoods based on appreciations of items. Afterwards the system generates recommendations with users in the neighborhood that an item would prefer.

Cosine Similarity

$$sim(u, v) = \frac{\sum_{i \in I_{uv}} r_{ui} r_{vi}}{\sqrt{\sum_{i \in I_u} r_{ui}^2} \sqrt{\sum_{i \in I_v} r_{vi}^2}} \cdot \frac{|I_{uv}|}{2} \quad (3)$$

Here, I_u is the set of all items with nonzero elements of user u in Q . I_v is the set of all items with nonzero elements of user v in Q .

Prediction

$$P(u, i) = \bar{r}_u + \frac{\sum_{v \in U} sim(u, v) (r_{vi} - \bar{r}_v)}{\sum_{v \in U} sim(u, v)} \quad (4)$$

Here, $\bar{r}_u$ is the average value of nonzero elements of user u in score matrix R . $\bar{r}_v$ is the average value of nonzero elements of user v in R .

ALGORITHM

Input: Rating

Output: Prediction for User

{

For each two users, do

Measure how similar they are in terms of having received similar ratings by similar users.

Similarity between users is measured as cosine between their ratings vectors.

Select n active users that have the highest similarity

Compute a prediction, based on an item for a user u is calculated.

Computing weighted sum of different users rating on item i

End for

}

D. RECOMMENDATION

Based on the similarity value we predict the user interest and recommend the TOP- N items recommendation. The recommender system utilizes data mining methods and prediction algorithms to prognosticate user's interest on items from massive amount of accessible items. Recommender systems make suggestions automatically for a particular user on the basis of the user's browsing history, ratings and comments given to an item. Recommender systems apply techniques and improve the recommendation task. Recommender systems commonly produce a list of suggestions in one of the two techniques through collaborative or content-based filtering. Content based algorithm recommender systems is the recommender system which work with profiles of users that are set up at the beginning. A profile has details of a user and the user's interest. Interest depends upon how user rated items.

IV. RESULTS AND DISCUSSIONS

This section discusses about the intermediate results of each component of the system and provides the performance evaluation of the implemented methodologies.

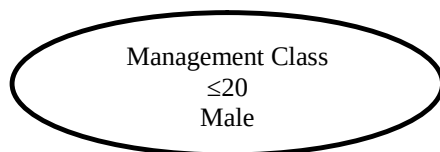
E.EXPERIMENTAL SETUP

In clustering firstwe given a set of users and then cluster the users based on attributes such as age, gender, occupation. Then apply the Euclidean distance.Based on the distance we form the cluster.

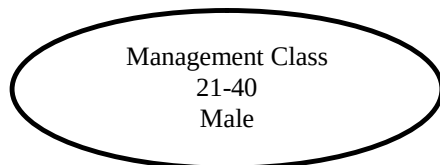
Table I

Clustering Results

Id	Age	Gender	Occupation
1	1	0	1
2	1	1	0
3	1	0	1
4	1	0	1
5	1	1	0



Cluster -1



ID	1	2	3	4	5	6	7	8	9	10
1	2	5	4	3	3	1	3	3	2	1
2	1	3	3	2	1	4	3	3	3	3
3	2	4	5	3	2	1	2	3	4	2
4	5	4	3	2	1	1	3	3	3	2
5	2	2	1	3	3	5	2	1	3	2
6	1	3	2	3	3	1	3	5	4	2
7	5	3	3	3	1	1	4	3	3	1
8	2	3	1	3	5	4	3	3	3	2
9	1	2	3	3	3	1	3	4	3	2
10	2	3	3	2	5	4	3	3	3	2

Similarly we form the cluster for remaining category.Totally 18 clusters are formed.

Table 1 show that the result clustered user.

Table II

Filling the Rating Matrix Results Based on Data Filling Approach

Table II show the result of filled rating matrix.In that rows represent the userid and column represent the columnid.User1 rated 2 to the Movie1.Similarly all the users rated the movies.

TableIII

User Based Collaborative Filtering Results

User-User similarity computation between user 3 and other users

Iid/ Uid	1	2	3	4	5	6	Cosine(i,5) (user-user)
1	5	4	5	4	5	4	0.889
2	4	5	2	4	3	4	0.938
3	1	3	2	2	2	3	0.465
4	1	2	2	3	2	4	0.372
5	0	3	3	1	1	0	1.0

The top-2 closest users to user 5 are users 1 and 2 according to both measures. Predictions are obtained for user 5 with respect to the unrated item.

Iid/ Uid	1	2	3	4	5	6
1	5	4	5	4	5	4
2	4	5	2	4	3	4
3	1	3	2	2	2	3
4	1	2	2	3	2	4
5	5	3	3	1	1	4

Table III shows sample result of User Based Collaborative filtering.

Table V

Recommendation Results

Userid	Movieid	Rating
1	7	-1
2	1029	-1

3	4	1
4	1172	1
5	3	-1

A recommendation is a list of items that the user will like the most. This interface is known as Top- N recommendation. So that mapped the value of rating 5 to 1 and below 5 are mapped to -1

User 1 is mostly interested on **Drama** category. So recommend the drama movie to the user 1

No	Movie ID	Title
1	131	Breakfast at Tiffany's
2150		Weekend at Bernie's
3175		Brazil
4 244		Smilla's Sense of Snow
5416		Fish Tank
6524		Temple Grandin

Table V shows sample results of recommendation.

F. PERFORMANCE ANALYSIS

To evaluate the performance of rating prediction we take on the widely used. Root Mean Squared Error and Mean Absolute Error for calculating the error between actual ratings and predicted rating value. The RMSE and MAE can be calculated as,

$$RMSE = \sqrt{\frac{1}{N} \sum_{i,j} (p_{i,j} - r_{i,j})^2} \quad (5)$$

$$MAE = \frac{\sum_{i,j} |p_{i,j} - r_{i,j}|}{n} \quad (6)$$

Where n is the number of total ratings

$p_{i,j}$ Is the prediction for user I for item J

$r_{i,j}$ Is the actual rating.

The Proposed Method RMSE and MAE Values are

0.978 and 0.782

IV CONCLUSION AND FUTUTRE WORK

Recommender systems play an important role in the providing of user-specific services by filtering the large variety of available data to draw out information on user

preferences. Nowadays, recommendation systems are increasingly gaining due to their large number of applications. The users cannot manage all the information available on Internet because of this is necessary some kind of filtering technique is needed.

Clustering technology can improve the scalability of collaborative filtering by reducing the nearest neighbor search space. User based CF techniques make great contributions to the rating prediction and recommendation. RS algorithms to mine user-item interaction data (such as rating) and provide useful recommendation lists for any target user. In our future work, a hybrid approach which consists of both model based and memory based algorithms in order to increase the performance of the system. A similar recommender system can be incorporated for Music or Book recommendation systems. It is vital that precise recommendations are provided to the users.

ACKNOWLEDGMENT

The authors wish to thank the Management and Principal of Mepco Schlenk Engineering College, for the support in carrying out this research work.

REFERENCES

- [1] C.X. Zhang, Z.K. Zhang, L. Yu, C. Liu, H. Liu, X.Y. Yan, Information filtering via collaborative user clustering modeling, Phys. A (2013)
- [2] B.M. Sarwar, Sparsity, Scalability, and Distribution in Recommender Systems, Uni-versity of Minnesota, 2001
- [3] C.L. Liao, S.J. Lee, A clustering based approach to improving the efficiency of col-laborative filtering recommendation, Electron. Commerce (2016).
- [4] C. Wang, L. Wang, J. Wang, S. Sun, C. Xia, Inferring the reputation enhances the co-operation in the public goods game on interdependent lattices, Appl. Math. Comput (2017)
- [5] H. Koohi, K. Kiani, A new method to find neighbor users that improves the per-formance of collaborative filtering, Expert Syst. Appl. 83 (C) (2017)
- [6] G. Linden, B. Smith, J. York. Amazon.com Recommendations, Item-to-Item Collaborative Filtering. IEEE Internet Computing, Volume 7, Issue 1, (2003).
- [7] Guoshuai Zhao, Xueming Qian and Xing Xie, Senior, " User-Service Rating Prediction by Exploring Social Users Rating Behaviors" IEEE Transaction on Multimedia, 2016.
- [8] M. D. Ekstrand, J. T. Riedl, J. A. Konstan. Collaborative Filtering Recommender Systems. Human-Computer Interaction, Volume 4, Issue 2, (2010).
- [9] Y. Koren, R. Bell, C. Volinsky. Matrix Factorization Techniques for Recommender Systems. Computer, Volume 42, Issue 8, (2009).
- [10] Y. Tian, J. Zhu, Learning from crowds in the presence of schools of thought, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, , 2012.
- [11] J. Gao, Y.W. Dong, M. Shang, S.M. Cai, T. Zhou, Group-based ranking method for online rating systems with spamming attacks, (2015).
- [12] Liji U, Yahui Chai, and Jianrui Chen, " Improved personalized recommendation based on user attributes clustering and score matrix filling" Science Direct -Computer Standards and Interfaces, July-2017

