2019 2nd Scientific Conference of Computer Sciences (SCCS), University of Technology - Iraq

A Proposed Model to Solve Cold Start Problem using Fuzzy User-Based Clustering

Nadia F. AL-Bakri
College of Science
Al Nahrain University
Baghdad, Iraq
nadiaf_1966@yahoo.com

Sukaina Hassan
dept. of Computer Science
University of Technology
Baghdad, Iraq
soukaena.hassan@yahoo.com

***Abstract*—In online environments, a vast amount of data is explored daily on the internet, such as news, movies, audios, and books. The interest of the target user is the greatest demand for recommender systems, and getting suitable information is a challenge. To develop a recommender system, collaborative filtering (CF) approach considers users who have similar ratings. Therefore, can compute the similarity of users when there are enough rated by users to items. A significant challenge of the collaborative filtering approach is cold start problem, which is how to make a recommendation for users who have few ratings than others. This work proposes a collaborative filtering model based on applying fuzzy c-means clustering on user's truthfulness information. A new fuzzy user-based similarity measure formula is suggested which combines user's rating with fuzzy truthfulness information using a combination coefficient. The experimental results using Movie Lens data set have shown an improvement in recommendation under sparsity and cold start condition.**

Keywords — Recommender System, Cold Start, Clustering, Fuzzy C-means, Pearson Correlation

I. INTRODUCTION

Recommender systems are tools that utilize the historical interest of a group of users to assist entities in that group to effectively explore new things of interest from a possibly tremendous set of choices. Collaborative filtering methods are mostly used in online item recommendation. Its mechanism is finding similarities among users using item ratings as a first step. Hence items can be recommended based on the similarity computation. Many similarity measures can be applied, such as Pearson's correlation and cosine similarity measure [1]. A popular problem in recommendations is the cold start problem. It refers to the significant degradation of recommendation quality when only a few numbers of interactions between target users and recommender system are available. In addition, finding nearest neighbors in user-item rating.

Matrix is achieved byusing the KNN collaborative filtering algorithm, which needs to search the whole user-item space to find the user's neighbors. Therefore, it will be difficult to meet the real-time recommendations due to the large number of users that affects the time-consuming. Response to these problems, a proposed model is presented in this paper to cope with these difficulties. This paper is prepared as follows: In section II, related works on this field is presented. In section III, an overview of recommender algorithms is shown. Clustering concepts and algorithm are subjected in section IV. The proposed work is presented in section V. Discussion on the experimental results is conducted in section VI. The last section is the conclusion of this work.

II. Related work

In what follows, some of the previous researches that are related to the techniques are used in collaborative filtering to overcome the cold-start problem is presented with employing different data sets.

Li, Q. and Kim B.M., 2003, [3] the authors presented an Item-based Clustering Hybrid Method (ICHM) that clusters item's features (content) to complement the user rating preferences. A formula is suggested, which is a linear combination of similarities between user-rating using Pearson measure and clustered rating using cosine measure. Experimental results on Movie Lens 100K dataset shown that this formulation was overcome the sparsity problem in data and improved the recommendation performance. In 2010 [4], the authors adopted a new hybrid model for solving Cold-start problem. Collaborative Filtering (CF), Content-Based (CB), Demographic-Based, the fusion of CF and CBF has been used as input approaches. The results are fused by using the optimistic weighted averaging (OWA) operator. Experimental results on MovieLens 100k dataset shown an increased performance compared with other methods. In 2013 [5], the authors adopted a method for solving Cold-start problem in collaborative filtering approach. Social sub-community division is implemented by using Pearson similarity measure and K-means clustering method to follow and user's history preferences and relationships between users. The ontology decision model is built by using Classification and Regression Tree (CART) on the basis of sub-community and user's static information to make a recommendation. Experiment on MovieLens 100k dataset shown that the suggested method had less MAE in comparison with other methods. In 2016[6], the authors presented a work that applies K-means clustering on learner's past educational data, parental information, and his current technical experience. The new learner is classified using KNN according to his attendance to an online c-test. Experiments on IMSAA online real dataset and MovieLens 100K datasets were shown a significant performance on cold start users. In 2018 [7], the author provided a new means to cope with cold users by the existing categories of items.

978-1-7281-0761-5/19/$31.00 ©2019 IEEE

Ranking of the suggested items is improved by calculating the average of each item's category and compared it with an average rating for the new user. The obtained results from conducting Movie Lens 100K, 1MB, 10MB dataset shows that in addition to increased processing speed, recall and precision have an acceptable improvement.

III. Recommender System Algorithms

Two most common recommender algorithms are: [8]

1) Collaborative filtering: items that users with similar preferences preferred previously will be recommended to the target user. The analysis of the user's features is not required.

~~2)~~ Content-based methods: these methods analyze item, or user features to explore items that areof particular interest to the user.

Both content-based and collaborative filtering have flaws in cold-start condition because users have no previous rating. So using both methods cannot perform accurately and as a result, no new items are recommended for cold user. Because of the cold start problem, the data will be sparse, and so the overall number of items in the system is more than the number of items that users have rated [9].

Two situations of the cold-start problem:

1) Cold-start users: It is a challenge caused when a new user has rated just a few items, and this leads to the weak similarity between users. Items of other users cannot be suggested based on most recommendation algorithm [8].
2) Cold-start items: It is a challenge caused by items which have no prior ratings, and therefore, they are not likely to be recommended [10].

Collaborative filtering (CF) techniques are categorized into memory-based CF and model-based CF.

According to the subject of this paper, memory-based collaborative filtering is considered.

A. Memory-Based Collaborative Filtering

This technique finds a list of suggested items for a target user based on similar users (user-based) or similar items (item-based). These methods have succeeded in a widespread achievement in existent life applications.

The proposed model conducted user-based method [11].

1) User-Based Collaborative Filtering

This technique is based on a comparison between user's ratings on thesame items by calculating the similarity between users, and then, computation is done to predict a rate for an item by the target user. The correlation of user *u* to user *v* is calculated by using the formula below, which is called the *Pearson correlation coefficient* for users:

$$Sim(u,v)=\frac{\sum_{i\in I}\left(r_{u,i}-\overline{r_u}\right)\left(r_{v,i}-\overline{r_v}\right)}{\sqrt{\sum_{i\in I}\left(r_{u,i}-\overline{r_u}\right)^2}\sqrt{\sum_{i\in I}\left(r_{v,i}-\overline{r_v}\right)^2}} \qquad (1)$$

Where i ∈ I, and I is overall co-rated items for both user u as well as user v given their rates. $r_{u,i}$means the user u rates for item i ,$\overline{r_u}$,$\overline{r_v}$are users u ,v mean rating[11].

2) User-BasedCollaborative Filtering Prediction

A prediction for a target user(a) on a specific item i is computed by summing the mean of the active user(a) with a summation of the rating's weights on that item i, and normalized by the sum of the weights. The user-based prediction formula is presented below:

$$\Pr edict(a,i)=\overline{r}_{a}+\frac{\sum_{u\in U}sim(a,u)\cdot\left(r_{u,i}-\overline{r}_{u}\right)}{\sum_{u\in U}\left|sim(a,u)\right|} \qquad (2)$$

u ∈ Uin which U specifies target user's neighbors (highest similarities).

Sim (a,u) is a set of similar users (u's) to target user (a)[11].

IV. Clustering Algorithms

The main objective of clustering methods is partitioning a given set of unordered items to obtain a number of clusters in which items within one cluster are similar. Clustering is considered a machine learning method, and it is unsupervised [12].

A. Fuzzy Clustering

The main objective of fuzzy clustering is to apply a clustering method with fuzziness. The fuzzy concept in clustering means that each object belongs to all constructed clusters with a certain degree of fuzzy membership. On the contrary, crisp clustering means that each object is assigned to a distinct cluster. When clusters are not well separated, this approach is particularly useful. Moreover, more relations can be discovered between the target object and the closest cluster when dealing with membership function. Fuzzy C Means (FCM) is one of the most popular fuzzy clustering algorithms which aim to find a fuzzy cluster with minimum cost function [13].

B. Fuzzy C-Means (FCM)

FCM approach is finding a partition (C fuzzy clusters) for a set of data points $x_j \in R^d, j = 1 \dots N$with minimal cost function [13]:

$$j\left(U,M\right)=\sum_{i=1}^{C}\sum_{j=1}^{N}u_{i,j}^{m}D_{i,j} \qquad (3)$$

where μ=[$\mu_{i,j}$] is the fuzzy membership value for cluster i with its object j.

C=[c_1,…,c_c] is the cluster mean (center) matrix.

$m \in [1,\infty)$is the fuzzification parameter, where m=2.

$D_{i,j} = D(x_j, m_i)$Isa value specifying the distance between x_j andm_i.

The following is the FCM algorithm [13]:

1) Choose values for *m and c* and a small positive numberε. Initialize the mean matrix C randomly. Assign step value t=0.

2) Computes(when t=0) or updates (t 0>) the membership matrix $\mu$by:

$$\mu_{i,j}^{(t+1)} = \frac{1}{\sum_{t=1}^{c}\left(\frac{D_{tj}}{D_{i,j}}\right)^{1/(1-m)}}$$

For i=1,…,c and j=1,…,N

3) Updates the mean matrix *C* according to the new values of$\mu_{i,j}$:

$$c_i^{(t+1)} = \frac{\sum_{j=1}^{N}(u_{i,j}^{(t+1)})^m x_j}{\sum_{j=1}^{N}(u_{i,j}^{(t+1)})^m}$$

4) Repeats step 2 and step 3 until minimum cost function is achieved $\left\|C^{(t+1)} - C^{(t)}\right\| < \varepsilon$

V. THE PROPOSEDWORK

The proposed model is divided into2 phases. Algorithms (1) and (2) is an outline of the proposed model.

```
Algorithm(1): Training Phase (Offline)
Input:    Receiving Movie Lens dataset
Output: Proposed Fuzzy User-Based Similarity Measure
Begin
Step1:  Loads "MovieLens data set" from Group Lens website [14]. // (investigated and analyzed to construct two matrices)
Step2:  Constructs user-movie rating matrix.
    // where the rows represent MovieLens training users and the column representing the movies. The corresponding intersection cells are filled with given training user's ratings to movies.  The ratings. range is between[1,5]
Step3:   Applies Pearson similarity measure on user-movie rating matrix to produce user-based similarity matrix.
Step4:   Constructs user's truthfulness matrix.
    // where the rows represent Movie Lens users and the columns representing the evaluated information for each user from user-movie rating matrix. The user's truthfulness matrix consists of three fields for each user by applying the following formulas:
        4.1 Compute User's activity for user (u)
        // which is the number of movies ($M_u$) the user (u) rated.
            User_activity (u) =count ($M_u$)
        4.2 Computes User_probity
        // which is the mean square deviation for user (u).
            User_propity(u)= $\sqrt{\frac{\sum_{j\in M}(R_{u,j}-\overline{R_u})}{count(M_u)}}$
        //   where $R_{u,j}$ is user (u) rating to movie j,$\overline{R_u}$is  average  ratingsfor user (u).
        4.3  Compute User_friends score
        // which is computed by taking the maximum average for user (u)'s nearest friends. The nearest friends can be found using K nearest neighbor method for the specified user (u) utilizing the constructed user based similarity matrix.
            User_friend score (u) =MAX (AVG ($u_i$)
        //Where i=1...K, K the number of nearest users.
Step5: Applies fuzzy c-means clustering on truthfulness matrix to cluster similar users with a certain degree of membership for each user. The member ship value is between [0, 1].
Step6: Computes the following proposed similarity formula which is the product of the two computed similarity matrices. The result is fuzzy-based matrix.
        $sim(i,j)\ fuzzy - based = sim_{user-based} \times (c) + sim_{fuzzy\ truthfullness} \times (1-c)$  …."Eq. 3"
    Where $sim(i,j)\ fuzzy - based$ is the similarity between user i and user j. $(c)$ is combination coefficient.
End
```

```
Algorithm(2): Testing Phase
Input: Target new User with Few Rating.
Output: Recommended Items.
Begin
Step1: randomly select users from test Movie Lens dataset with few rating   //consider them cold start users.
Step2: Computes the truthfulness information for new users.
Step3: Applies Pearson correlation coefficient on user-rating matrix with the new users.
Step4: Re-cluster the truthfulness matrix with the new users.
Step5: Applies the proposed similarity formula "Eq. 3".
Step6: Applies user-based prediction formula "Eq. 2" to predict the unrated movies for the target new user.
Step7: recommend the highest 10 estimated rating for movies found from step4 and suggest them to target new user.
End
```

VI. DISCUSSION

The proposed model is confirmed byusing the Movie Lens data set. It is used to receive the user's rating for movies and further to be analyzed and preprocessed. Its data consists of 100000 ratings from 943 users and 1682 movies. An average of about 20 rates for each user is given. In this model; users with a few rating are selected randomly to represent cold start users. As an example, user 4 and user 9 have no common rating and have zero ratings for the first 10 movies Table I, but they have rated 12 and 13 movies respectively (as shown in Table III in the whole dataset. Also, the similarity between them is zero (Table II).The preprocessing is achievedoffline to reduce the time needed for recommendation. After applying the proposed model offline, as shown in algorithm (1), a user-movie rating matrix is constructed as shown in Table I. Pearson correlation is applied on user-movie rating matrix. Its value ranges [-1, 1] as depicted in Table II.

Computation of the truthfulness for each trained user is achieved as depicted in Table III, which describes the user's behavior. The reason behind clustering the user's truthfulness information is to build a fuzzy matrix where each user has a memberships belonging to each cluster as depicted in Table IV. These membership values will be used in similarity computation between users. Then a modified formula "Eq. 3" is applied. As a result, this process will alleviate the sparsity problem that leads to accurate similarity values and especially with cold start problem. Here in this work, fuzzy c-means is implemented with two clusters. For example, user1 belongs to cluster1 with a membership 33% and cluster2 with a membership 66%. Table V depicts the center of the two clusters. Comparing this work with previous ones in [3], the authors used item's information instead of user's information and similarity computation is on item-based CF.Each user has a membership value byusing a fuzzy set. In [6], authors defeatcold start by using an asked questionnaire process and user's demographic information.In [7], items are categorized according to their movie genre(movie type), and a comparison is achievedwith the target user average rating. All these papers did not use information extracted from the user's rating as in this work. In[5],the authors utilized a decision tree in addition to clustering to make recommendations.

In user-based collaborative filtering, the prediction process is based only on user-based similarity matrix (as shown in Table II. It can't make predictions for users with no common rating. As an example, user 4 with user 6 and user 9 have similarity value equal to zero. However, in the proposed model, a similarity and prediction for new users are based on the modified equation "Eq. 3" which takes a user's truthfulness into consideration. The combination coefficient (c) equals 0.4 according to trial and error process. This coefficient is used as a biased value.

VII. CONCLUSION

In traditional collaborative filtering, it is hard to use it in a recommendation for a new user since there is no previous history (rating) for him. However, figuring the truthfulness of users in combination with the user-movie rating matrix in the proposed model will influence in solving the cold start problem. Using fuzzy c-means clustering, have a great impact on prediction performance. From the description of the proposed model, it is observed that the proposed framework can fully recognize the strengths of fuzzy user-based filtering, alleviating, and solving the new user problem. This formulation overcomes the sparseness problem in data and improved the recommendation performance in cold start problem.

TABLE I. SAMPLE OF CONSTRUCTED USER-MOVIE RATING MATRIX FROM MOVIELENS DATASET

	Movie1	**Movie2**	**Movie3**	**Movie4**	**Movie5**	**Movie6**	**Movie7**	**Movie8**	**Movie9**	**Movie10**
User 1	5	3	4	3	3	5	4	1	5	3
User 2	4	0	0	0	0	0	0	0	0	2
User 3	0	0	0	0	0	0	0	0	0	0
User 4	0	0	0	0	0	0	0	0	0	0
User 5	0	0	0	0	0	0	0	0	0	0
User 6	4	0	0	0	0	0	2	4	4	0
User 7	0	0	0	5	0	0	5	5	5	4
User 8	0	0	0	0	0	0	3	0	0	0
User 9	0	0	0	0	0	0	4	0	0	0
User 10	4	0	0	4	0	0	0	0	4	0

TABLE II. USER-BASED SIMILARITY MATRIX

	User1	**User2**	**User3**	**User4**	**User5**	**User6**	**User7**	**User8**	**User9**	**User10**
User1	1.0000	0.9545	0.8555	0.9318	0.9285	0.9527	0.9401	0.9754	0.9690	0.9677
User2	0.9545	1.0000	0.9522	0.9918	0.9829	0.9565	0.9624	0.9664	0.8907	0.9770
User3	0.8555	0.9522	1.0000	0.9484	1.0000	0.8808	0.8721	0.8785	0.0000	0.9214
User4	0.9318	0.9918	0.9484	1.0000	1.0000	0.0000	0.9058	0.9816	0.0000	1.0000
User5	0.9285	0.9829	1.0000	1.0000	1.0000	0.9355	0.9036	0.9537	0.8807	0.9340
User6	0.9527	0.9565	0.8808	0.0000	0.9355	1.0000	0.9579	0.9885	0.9583	0.9796
User7	0.9401	0.9624	0.8721	0.9058	0.9036	0.9579	1.0000	0.9645	0.9337	0.9772
User8	0.9754	0.9664	0.8785	0.9816	0.9537	0.9885	0.9645	1.0000	1.0000	0.9839
User9	0.9690	0.8907	0.0000	0.0000	0.8807	0.9583	0.9337	1.0000	1.0000	0.9931
User10	0.9677	0.9770	0.9214	1.0000	0.9340	0.9796	0.9772	0.9839	0.9931	1.0000

TABLE III. THE TRUTHFULNESS MATRIX FOR MOVIELENS USERS

	User_activity	User_probity	User_friend_score
User1	262	1.45	4
User2	52	1.71	4
User 3	44	1.66	4
User 4	13	0.33	3
User 5	164	2.45	4
User 6	201	1.33	5
User 7	393	1.77	3
User 8	49	2.42	3
User 9	12	0.22	2
User10	174	1.47	5
User13	626	1.37	3

TABLE IV. MEMBERSHIP FOR 11 USERS AFTER CLUSTERING TRUTHFULNESS INFORMATION

	User 1	User 2	User 3	User 4	User 5	User 6	User 7	User 8	User 9	User 10	User13
Cluster 1	0.3345	0.0095	0.0128	0.0282	0.0406	0.1128	0.8913	0.0107	0.0288	0.0558	0.9441
Cluster 2	0.6655	0.9905	0.9872	0.9718	0.9594	0.8872	0.1087	0.9893	0.9712	0.9442	0.0559

TABLE V. CLUSTER1 AND CLUSTER2 CENTERS

Cluster1	496.88	1.550	3.07
Cluser2	95.49	1.65	3.72

REFERENCES

[1] A. Abdelwahab, H. Sekiya, I. Matsuba, Y. Horiuchi, and S. Kuroiwa, "Collaborative filtering based on an iterative prediction method to alleviate the sparsity problem." ACM, Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services [internet].; pp: 375-379, December, 2009. DOI:10.1145/1806338.1806406.

[2] Z. Wu, Y. Chen, and T. Li, "Personalized recommendation based on the improved similarity and fuzzy clustering." In *Information Science Electronics and Electrical Engineering (ISEEE), 2014 International Conference on* IEEE. Vol 2. pp. 1353-1357. April.2014.

[3] Q.Li, and B.M. Kim, "Clustering approach for hybrid recommender system." In null (p. 33). IEEE. October.2003.

[4] J. Basiri, A. Shakery, B. Moshiri, and M.Z. Hayat, "Alleviating the cold-start problem of recommender systems using a new hybrid approach." In Telecommunications (IST), 2010 5th International Symposium on pp. 962-967. IEEE. December.2010.

[5] M. Chen, C. Yang, J. Chen, and P. Yi, "A method to solve cold-start problem in recommendation system based on social network sub-community and ontology decision model." In 3rd International conference on multimedia technology ICMT. 2013.

[6] G. Sakarkar, and S.P. Deshpande, "Clustering based approach to overcome cold start problem in intelligent e-learning system." International journal of latest trends in engineering and technology (IJLTET).7(1).2016.DOI:10.21172/1.71.001.

[7] H. Jazayeriy, S. Mohammadi, and S. Shamshirband, "A Fast Recommender System for Cold User Using Categorized Items." Mathematical and Computational Applications, 23(1), p.1.2018.

[8] V.A. Rohani, Z.M. Kasirun, S. Kumar, and S. Shamshirband, "An effective recommender algorithm for cold-start problem in academic social networks." Mathematical Problems in Engineering. 2014.

[9] Z. Xu, and Q. Fuqiang, "Collaborative filtering recommendation model based on user's credibility clustering." In Distributed Computing and Applications to Business, Engineering and Science (DCABES). 2014 13th International Symposium. pp. 234-238. IEEE. November.2014.

[10] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, "Recommender systems survey." Knowledge-based systems, 46, pp.109-132.2013.

[11] N.F. Al-Bakri, and S.H. Hashim, "Reducing Data Sparsity in Recommender Systems." Journal of Al-Nahrain University-Science, 21(2), pp.138-147.2018.

[12] R. Kruse, C. Döring M.J. Lesot, "Fundamentals of fuzzy clustering. Advances in fuzzy clustering and its applications." 2007 Jun 13:3-0.

[13] R. Xu, and D.Wunsch, "Survey of clustering algorithms." IEEE Transactions on neural networks. 16 (3):645-78. May, 2005.

[14] https://grouplens.org/datasets/movielens.