Hybrid Recommender System based on Fuzzy Clustering and Collaborative Filtering

Sumit Kumar Verma
Department of Computer Engineering
Malaviya National Institute of Technology, Jaipur
sumitkverma03@gmail.com

Namita Mittal
Department of Computer Engineering
Malaviya National Institute of Technology, Jaipur
nmittal@mnit.ac.in

Basant Agarwal
Department of Computer Engineering
Malaviya National Institute of Technology, Jaipur
thebasant@gmail.com

Abstract— Recommender systems have achieved widespread success for e-commerce companies. Significant growth of customers and products poses key challenges for recommender system namely sparsity and scalability. In this paper, a hybrid system is proposed that is capable of handling these issues that is based on collaborative filtering and fuzzy c-means clustering algorithms. Experimental results show the effectiveness of the proposed recommender system.

Keywords— Recommender System, Collaborative Filtering, Fuzzy Clustering (FCM).

I. INTRODUCTION

Nowadays recommender systems have evolved tremendously due to increasingly growth of Internet. E-commerce companies are using recommender systems to give suggestions to purchase the items which their customers may likely be interested in. Recommender system has been proved to be excellent in the sales of the e-commerce companies. In recent years, it is common for e-commerce Businesses to implements this recommender system. These systems recommend items such as Books, CDs/DVDs , movies, music etc to their customers. Examples of recommender system are NetFlix which recommend movies to the users and Amazon which recommend the books that their users may like. There are several different ways to make recommendations, including providing top list of items, making suggestion based on demographic data and making recommendation by analyzing past user interaction of the user with the system. Among all, Collaborative Filtering (CF) is one of the best techniques proposed by [Goldberg et al., 1992]. Basically Recommender System can be classified in three types, Collaborative Filtering, Content-based recommender system and Hybrid approach. Content based recommender system use the information about the item or user's profile. So it's quite challenging to get the information if the item is multimedia product like video, music.

In Collaborative Filtering Approach, the key element is past user interaction with the System. CF recommender system uses the known ratings of the items made by the users to predict rating of new user-item pairs. The idea behind this is that two users most probably continue liking similar items if they have already liked similar ones. Here, focus on the Collaborative Filtering Approach in this paper.

In the proposed approach,initially clusters of the item on the basis of profile of the item are created using fuzzy c-means clustering algorithm. Then, these clusters use item based collaborative filtering approach to predict the rating.

Related work is described in Section II. Section III discussthe proposed approach in detail. Experiment results are discussed inSection IV. And finally, conclusion is presented in Section-V.

II. RELATED WORK

In [3],authors proposed an algorithm of collaborative filtering which is based on the difference between the users and items. First they compare all different approaches of collaborative filtering. According to their approach, they were not consider the relation between items and users, other than they consider the difference between them. On the condition that there are some users, who inclined to give positive rating, leaving negative ratings for really bad items, while other user, save their highest rating s for the best item and tends to give negative ratings. So according to their approach, first find the tendencies of items and users and on the basis of this, recommendation is done.

In [11],authors proposed a hybrid approach of recommender system, which takes advantages of both content-based and collaborative filtering technique for recommendation. According to their approach they first find the similar user with the help of k-means clustering algorithm then find the content that the users of the same cluster rated high. Add this content in the list of contents and then apply fuzzy c-mean algorithm to find contents in the same cluster as that of the content requested. And in the end find common set from precious two computed set. That is the best result for the search.

In [15],authors proposed a composite collaborative filtering algorithm for personalized recommendation. According to their approach , they combine user-based CF and item-based CF together, and using a spearman rank correlation coefficient instead of pearson correlation to ensure the equal space in logic area of the data, which do not need to be receive in pairs from the normal distribution. According to this algorithm, CF algorithm predicts the similar item set by the data has been

provided, and then use second algorithm to get the final assessment and solve the problem called singular data.

III. PROPOSEDAPPROACH

Approach is divided into two phase

Phase 1: Clustering: of items on the basis of item's profile

Phase 2: Item-Based Collaborative Filtering: Apply item-based collaborative filtering algorithm on each cluster for the predicting the rating.

Overview of our approach

Step 1: Clustering is done on the basis of item profile.

Step 2: Apply item based collaborative filtering approach on each cluster to predict the missing rating in the user-item matrix.

Step 3: To reduce the Cold Start problem. *New User*: Rating must be given on the specific (threshold) number of items to get recommendation.

New Item (NI) : Rating of new item by user U is given by

Rating of NI(U) = Average rating of the user U within that cluster.

Step 4: To reduce the scalability problem :

Clustering is pre-processing step. This algorithm is run periodically or triggered after some particular (threshold) number of new items add in the system.

Phase 1: Clustering

A. K-mean Clustering

Overview of this approach

K-mean is apply on the dataset the item's profile. In this paper use Movielensdataset(http://www.movielens.com). This dataset consists of 1,00,000 ratings (1-5) from 943 users on 1682 movies. This dataset also consist profile of each movies, that to which genre(ex: Comedy, Action, etc) the movie(item) is belongs to. In the dataset , there are 19 different genre. And on the basis of these genre , clustering is done. Movies or items are divided into k cluster. And on these k cluster apply item-based collaborative filtering.

Select random k number of items from the item dataset and use them like k center of the clusters. Second step is to calculate distance between all the item to all the cluster's center.

Assume 10 movies , and each movie profile is describe by 3 genre(i.e. comedy, action, musical) see in table 1.

	Comedy	Action	Musical
M1	0	1	0
M2	1	0	1
M3	1	1	0
M4	0	1	1
M5	0	0	1
M6	1	0	0
M7	0	0	1
M8	1	1	1
M9	1	1	0
M10	1	0	1

Table 1

Use Manhattan distance function, to find dissimilarity(or similarity) between the objects(movie).

$$d(ij,)=|x_{il}-x_{jl}|+|x_{i2}-x_{j2}|+...+|x_{ip}-x_{jp}| \quad (1)$$

After applied K-means clustering, movies grouped into 2 clusters, as shown in table 2.

Cluster No.	Cluster Set
Cluster 1	{M1,M4 }
Cluster 2	{M2,M5,M7,M10}
Cluster 3	{M3, M6, M8, M9}

Now, on all of these 3 clusters, Item-based collaborative filtering is apply in Phase 2.

B. Fuzzy C-Mean Clustering

Fuzzy C-Mean is well known Fuzzy Clustering Algorithm. This allow one item belong to two or more clusters with a membership. It is based on minimization of the following objective function :

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\left\|x_i - c_j\right\|^2$$

'm' is any real number greater than 1, N is number of item, C is number of cluster, u_{ij} is the degree of membership of x_i in the cluster j.$x_i$is the ith of d-dimensional measured data. And c_j is the d-dimension center of the cluster.

Apply FCM on the same example of table 1. Results are shown in table 3

	Cluster 1	Cluster 2	Cluster 3
M1	0.81	0.03	0.16
M2	0.01	0.97	0.02
M3	0	0	1
M4	0.97	0.02	0.01
M5	0.08	0.91	0.02
M6	0.07	0.21	0.72
M7	0.08	0.91	0.2
M8	0.4	0.22	0.38
M9	0	0	1
M10	0.01	0.97	0.02

Table 3

Table 3 give the membership of each item to each cluster. Sum of all the membership of each item is equal to 1.Assume threshold value for acceptance of item in to some cluster. If membership of any item is greater than equal to this threshold value then, that item is the member of that cluster. By this ,clusters are formed having possibility of one item may part of more than one cluster.

Let assume the threshold value *0.15*. Clusters formed like that

Cluster No.	Cluster Set
Cluster 1	{M1,M4,M8 }
Cluster 2	{M2,M5,M6,M7,M8,M10}
Cluster 3	{M1,M3, M6,M7, M8, M9}

C. To reduce Cold Start Problem (FCM Approach)

When new item or new user is add into the system, then there is problem to make recommendation including these new items or new user, as they are not yet rate any item if new user is add. And they are not yet get any rating ,if new item is add. This issue is known as cold start problem. To reduce this

a) if new user add in the system: New user are bound to give rating on the specific(threshold) number of items to get recommendation.

b) if new item add in the system: rating of new item M, by user U is given by:

$$Rating\ of\ M(U) = \sum_{c=1}^{n} \overline{r_c(U)} * \mu_c(M) \quad ()$$

Where$\overline{r_c(U)}$is average of user U in cluster c. And $\mu_c(M)$is membership value of new item M to cluster c

Phase 2: Item Based collaborative filtering

For this phase, output of Phase 1 (Clustering) works as an input. Output of Clustering phase is item clusters. By this, get rating matrix of each cluster. Apply item-based collaborative filtering over each cluster.

Item-based collaborative filtering is one of the popular collaborative filtering approach. When apply user-based collaborative filtering approach to the millions of users and items, it do not scale well, because of the computational complexity of the search of the similar user. So in 2001 Sarwar et al. proposed a new approach which works on the similar item rather than the similar user.

In this approach first the similarity between pairs of item *i* and *j* are computed offline using Pearson correlation formula, so that we get that matrix of *n* * *n*having the similarity between each pair of items. And the Pearson correlation formula is given by :

$$S_{i,j} = \frac{\sum_{u \in U}(r_{u,i} - \bar{r}_i)(r_{u,j} - \bar{r}_j)}{\sqrt{\sum_{u \in U}(r_{u,i} - \bar{r}_i)^2}\sqrt{\sum_{u \in U}(r_{u,j} - \bar{r}_j)^2}}$$

Where *U*is the set of all users who have rated both the item *i* and *j*.$r_{u,i}$ rating of user *u* to item *i*,$\bar{r}_i$is the average rating of the item *i* across all the users.

Now we have the similarity matrix, having the similarity between each pair of items. By the help of this matrix, find the most similar *k*item to calculate the prediction. So the rating of item *i* by active user *a*can be predicted by using the simple rating average.

$$P_{a,i} = \frac{\sum_{j \in K}(r_{a,j} * S_{i,j})}{\sum_{j \in K}|S_{i,j}|}$$

Where K is the set of neighbor of *k* items rated by active user *a* and most similar to item *i*.

Now, for our scenario, we have k cluster of items and on the basis of these clusters, we made k item- user matrixes. And on these k item-user matrixes we apply this item- based collaborative filtering algorithm individually. And get the prediction rating matrix of all users to all items.

But there is one more issue encounter while implementing the second approach i.e. by FCM clustering. Issue is that, as the FCM clustering allow one item comes in one or more cluster with some membership. So there comes situation when one item comes in more than one cluster and its prediction is done on each cluster. So to solve this problem, we provide a rule that prediction rating of item *i* is the average of prediction rating of all cluster given by

$$P_{a,i} = \frac{\sum_{c \in C} P_{a,i}^{c}}{N_C}$$

Where *C* is the set of cluster which predicted the rating of active user *a* on item *i*,$P_{a,i}^{c}$ is the predicted rating of user *a* on item *i* according to cluster *c*. And N_C is number of clusters which predicted the rating of item *i* by user *a*.

IV. Experiment Results

In this section, we describe the dataset and methodology for the comparison between traditional and proposed collaborative filtering approach, and present the results of our experiments.

A. Data Set

We use MovieLens collaborative filtering data set to evaluate the performance of proposed approach. This data sets were collected by the GroupLens Research Projectat the University of Minnesota. This site is now has over 45000 users who have expressed opinions on 6600 different movies. We randomly selected enough users to obtain 100,000 rating from approx. 1000 users on 1680 movies with every user having at least 20 ratings and also having the simple movie profile information in form of movie belong which genre. There are 19 different genre. The rating are on a numeric five point scale from 1 to 5.

B. Evaluation Metrices

Several matriceshave been proposed for assessing the accuracy of collaborative filtering methods. They are divided mainly into two categories: statistical accuracy metrics and decision-support accuracy metrics. In this Paper, we use the statistical accuracy metrics.

Statistical accuracy metric evaluates the accuracy of a prediction algorithm by comparing the numerical deviation

of the predicted ratings from the respective actual user ratings. Some of them are MAE(Mean Absolute Error), RMSE(Root Mean Square Error). Both these were computed on the result data and provided the same conclusions.

If 'n' is the number of actual ratings in an item set, then MAE is defined as the average absolute difference between the n pairs. Assume that p1,p2,p3….pn is the prediction rating as a output of prediction algorithm and corresponding real ratings data set of users is q1,q2,q3….qn. MAE is defined as:

$$MAE = \frac{\sum_{i=1}^{n} |p_i - q_i|}{n}$$

The lower the MAE, the more accurate the predictions would be.MAE has been computed for different prediction algorithms.

Result Table 1 show the MAE and RMSE of Linear Fuzzy Clustering Based CF and Our Approach 1 and Approach 2. It is show that MAE of our approach based on FCM is better than all. As it first make cluster on the basis of items profile, So the similarity is to be computed between the similar items. That is make it better.

We Apply our both the approach on 5 disjoint datasets of movielens, and the result shown in Figure 1 in form of MAE. In the figure it is shown that our Fuzzy C-means based approach is better that k-means based approach. The reason of this is that, K-means is a hard clustering approach so there is the constraint that one item is only belong to any of the one cluster. But it is not true in real life scenario , there may be is chance that one item belongs to more than one cluster. For that reason, in our second approach we use Fuzzy C-means clustering (soft clustering), and get better results.

Apply our approaches on 3 different datasets which shown in figure 2 , which differentiate by rating density. x = 0.2 means there are 20,000 rating in the dataset. Similarly x = 0.5 means 50,000 ratings. Result shows that MAE of Fuzzy C-means Based Approach is better than K-means based approach.

Result Table1: MAE and RMSE of Different CF algorithms

Algorithm	MAE	RMSE
Linear FCM based CF	0.7160	
Our first approach (k-means)	0.7286	1.0101
Our second approach(fcm)	0.6806	0.9459

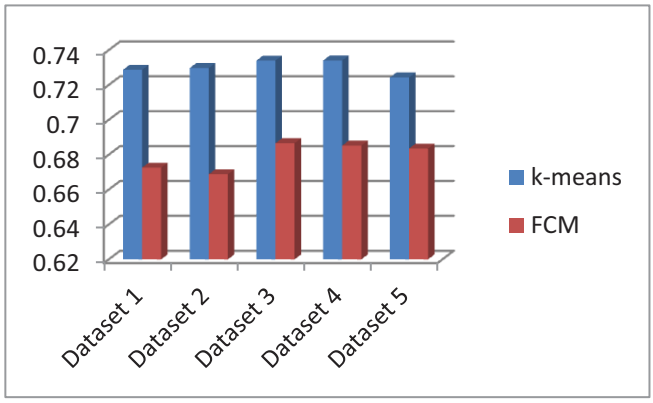

Figure 1 Mean Absolute Error corresponding to 5 disjoint dataset

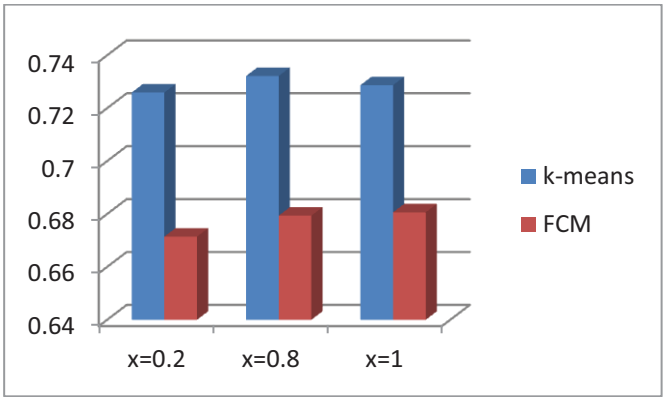

Figure 2 Mean Absolute Error corresponce to different proportion of dataset

V. Conclusion

In this paper, a hybrid system is proposed that is capable of handling sparsity and cold start problem by using collaborative filtering and fuzzy c-means clustering algorithms. Experimental results show the effectiveness of the proposed recommender system. Proposed system shows the effectiveness of the fuzzy clustering algorithm over k-means method for recommender system.

Proposed approach handles the Cold- start problem very efficiently. No requirement of pre-processing step (i.e. clustering) while new user adds into the system. It is require only when new item is adds in the system, as it works on item profile.

Proposed approach is memory-based approach , so there is no time waste in training the model.

References

[1] Prem Melville , Vikas Sindhwani, Recommender System,Encyclopedia of machine learning,2010.

[2] Robert M. Bell and Yehuda Koren, improved neighborhood-based Collaborative Filtering, KDD cup' 07, ACM, 2007.

[3] Fidel Cacheda, Victor Carneiro, Diego FernanDEZ, Tendencies based CF approach for Recommendation, ACM transaction on the web, vol 5, no. 1, Feb,2011.

[4] Huseyin Polat,Wenliang Du, SVD based CF approach for Recommendation, SAC'05, ACM, 2005.

[5] Bamshad Mobasher, Robin Burke, JJ Sandvig, Model-based CF approach against profile injection attacks, American Association of Artificial Intelligence,2006 Feb,2011.

[6] J. Ben Schafer, Dan Frankowski, Jon Herlocker, Shilad Sen,Collaborative Filtering Recommender systems,The Adaptive Web, LNCS 4321, Springer-verlag Berlin Heidelberg 2007.

[7] Jinlong wu, Tiejun Li,Modi_ed Fuzzy C-Means Algorithm for Collaborative Filtering,2nd-netix-KDD workshop,ACM, august 2008.

[8] Namita Mittal, M.C. Govil, Richi Nayak, Hybrid Clustering Based Filtering Approach,IMECS 2008.

[9] Liang Hu, Webo Wang, Feng Wang, Xiaolu Zhang, Kuo Zhao,Composite CF algorithms for personalized recommendation,Journal of Software, Vol. 7,No. 9, Academy Publication, 2012

[10] Mohammad Khabaaz, Laks V.S.Lakshmanan,Top-k Algorithms for item-based Collaborative Filtering, EDBT 2011, ACM 2011.

[11] Qilin Li, Mingtian Zhou,E_cient Collaborative Filtering Prediction algorithm,IEEE 2003.

[12] Prem Melville, Raymond J. Mooney, Ramadass Nagarajan,Content-boosted Filtering for improved recommendation,eighteenth National conference on arti_cial Intelligence,Canada, July 2002.

[13] SongJie Gong,A CF Algorithm based on user clustering and item clustering, Journal of software, vol. 5, No. 7, July 2010.

[14] , D. Nepolean, P. Ganga Laxmi, An efficient K-means clustering Algorithm for reducing time complexity using uniform distribution data points, IEEE 2010.

[15] , Jiawei Han and Micheline Kamber,Data Mining Concepts and techniques ,Second Edition, New Delhi, ISBN:978-81-312-0535-8.2006.

[16] K.L. Wu, M.S. Yang, Alternative C-means clustering algorithm, Pattern Recognition ,pp 2267-2278,2002