Collaborative Filtering Recommendation Model Based on k-means Clustering

Nadia Fadhil AL-Bakri [1*] and Soukaena Hassan Hashim [2]
[1] Department of Computer Science, AL Nahrain University, Baghdad-Iraq.
[2] Department of Computer Science, University of Technology, Baghdad-Iraq.
* Corresponding author: nfi@sc.nahrainuniv.edu.iq

Abstract

In this age of information load, it becomes a herculean task for user to get the relevant things from vast number of information. This huge number of data demand specially designed Recommender system that can plays an important role in suggesting relevant information preferred by the users. From this point, this paper presents a modest approach to enhance prediction in MovieLens dataset with high scalability by applying user-based collaborative filtering methods on clustered data. The proposal consists of three consequence phases: preprocessing phase, similarity phase, prediction phase. The experimental results obtained conducting K-means clustering and correlation coefficient similarity measures against MovieLens datasets lead to an increase in the scalability of recommender system. [DOI: 10.22401/ANJS.22.1.10]

Keywords: K-means, recommender system, clustering, movies, Collaborative filtering.

1. Introduction

Internet has turn out to be an important part of social life, there is an exponential growth in the amount of electronic information and on-line facilities available, this has led to the problem of overloading extensive information - where required information are difficult to be retrieved by people from the huge set of choices. Today the problem is how to choose the right item out of huge amount of information and not about how to acquire accurate facts to make a decision. Many applications have been revealed and recommender system is one of them to fulfil the demand of intelligent data analysis [1].

This paper is concentrated on clustering user-based collaborative filtering against MovieLens user-movie rating dataset. The paper is prepared as follows: the related works are presented in section 2, on section 3 a passing background in recommendation process and memory-based collaborative filtering algorithm and its prediction approach is described, in section 4 clustering concepts are presented, the proposed model is presented in section 5, the experimental results in section 6 and to end with the conclusion.

2. Related Work

In what follows, some of the previous research literatures related to clustering techniques used in a combination with memory-based collaborative filtering are presented.

1- [2] Liao, Q. et al., in this paper, K-means clustering is improved using parallelism algorithm implementing MapReduce concept. The K-means performance is enhanced by reducing the repetition numbers and fastening processing speed per iteration. Two strategies are proposed for distance measure and for initial selection of centroids that is reliable with the scattering of the data. The experimental results shown that, more stability is achieved with the proposed method than the traditional K-means.

2- [1] Khurana, P. and Parveen, S., in this paper, a proposal of recommender system that uses hybrid approach using improved K-means clustering with Spearman's rank correlation similarity to recommend items to users. Experiments were conducted on the MovieLens dataset using MATLAB tool. A comparison is done between the traditional K-means vs. the proposed K-means, it was shown that the proposed method had RMSE 1.6850 whereas the traditional method had 1.7220. The quality of clustering using the improved method is greatly dependent on the selection of initial centroids and yielding an effective reduction in RMSE and time complexity.

3- [3] Kumar, M., et al., in this paper, a movie recommendation (MOVREC) is introduced that permits a user for selections from a specific attributes and a

recommendation is aggregated weight of diverse features by using K-means algorithm. It is based on collaborative filtering approach. This system is implemented in PHP and Apache Server.

4- [4] Zebin Wu.et al. In This paper, a personalized recommendation algorithm is proposed based on improved similarity and fuzzy clustering. The proposed similarity method was used to solve the inaccuracy caused by the sparseness and to find nearest neighbors. Then a recommendation is produced after clustering. By the experimental results, it was shown that with the increase of the number of nearest neighbors for target users, the proposed method show a trend of gradual decline and have MAE error about (0.80 to 0.85) compared with the traditional user collaborative filtering which has MAE error about (0.95 to 1). Therefore using the proposed method, the accuracy of the recommendation is improved efficiently and the real-time response speed was improved as well.

5- [5] Yao, G, et al., in this paper, a recommendation system is built based on : item-based and user-based.. The Amazon datasets (Gourmet Foods.txt.gs) was used. The pre-processing on the features of the customers is to create a two-key hash map by using the first key review/userId and the second key is product/productId for the user-based model. Also product/productId as the first key and review/userId as the second key for the item-base model. The values for both models are from the review/score. A similarity and prediction computation is calculated. Content analysis to target products with a full attributes is analyzed using content-based filtering (CBF). A Matching process is computed between the product's profile and user's profile. Then the products with high correlation with the user's profile are recommended.

3. Recommender System

Recommender system is an Information Retrieval application that assists customers discovering interested items from a vast collection of things.it recommends everything from movies, news, books, songs and Web sites to more complicated suggestions for electronic gadgets, matrimonial matches, and financial services. Recommendation algorithms generally categorize into [1]:

1- Collaborative filtering.
2- Content-based filtering.
3- Demographics-based filtering.
4- Hybrid approaches.

In this paper, collaborative filtering of memory based is used for calculation.

Collaborative filtering is a process of calculating similarities between user preferences then a group called neighborhood is built and a user gets recommendations to items based on his neighborhood.

Collaborative filtering (CF) techniques are categories into: Memory-based CF and model-based CF.

3.1 Memory Based Collaborative Filtering

This technique finds a list of suggested items for a target user based on similar users (user-based) or similar item (item-based). These methods have succeeded a widespread achievement in existent life applications .the proposed model conducted user-based method [6].

3.1.1 User-Based Collaborative Filtering

This technique is based on a comparison between user's ratings on same items by calculating a similarity *using Pearson correlation* between users, and then a computation is done to predict a rate for an item by the target user. The range of the similarity measure is between [-1,1]. The correlation of user u to user v is calculated using the formula below which is called the *Pearson correlation* for users [6]:

$$Sim(u,v) = \frac{\sum_{i \in I} (r_{u,i} - \overline{r_u})(r_{v,i} - \overline{r_v})}{\sqrt{\sum_{i \in I} (r_{u,i} - \overline{r_u})^2} \sqrt{\sum_{i \in I} (r_{v,i} - \overline{r_v})^2}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (Eq.1)$$

$i \in I$, I is overall co-rated items for both user u as well as user v given their rates. $r_{u,i}$ Is the rate for user u to item i and $\overline{r_u}$, $\overline{r_v}$ are users u ,v mean rating.

3.1.2 User-Based Collaborative Filtering Prediction

A prediction for a target user(a) on a target item i is computed by summing the mean of active user(a) with 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 in the following formula [6]:

$$\boldsymbol{predict}(\boldsymbol{a},\boldsymbol{i}) = \overline{\boldsymbol{r_a}} + \frac{\sum_{u\in U} \boldsymbol{sim}\,(\boldsymbol{a},\boldsymbol{u}).\left(\boldsymbol{r_{u,i}} - \overline{\boldsymbol{r_u}}\right)}{\sum_{u\in U} |\boldsymbol{sim}\,(\boldsymbol{a},\boldsymbol{u})|} \quad \text{.....(Eq. 2)}$$

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

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

$r_{u,i}$ means user (u) rate to item i.

4. Clustering Techniques

Clustering is an unsupervised classification technique in pattern recognition and data mining. Large number of applications uses it. In clustering, an unlabeled objects is defined as vectors in a multidimensional space, are gathered into groups in such a way that objects within one cluster are similar according to some conditions and dissimilar in different clusters. In clustering, the input space is partitioned into *K* areas depending on some similarity/dissimilarity metric, where *K* may or may not be given a priori. A similarity measure has to be defined based on which cluster assignments can be done Different type of clustering is used for recommender system like K-means, fuzzy C-mean, chameleon and hierarchical [7].

4.1 The Traditional K-Means Algorithm

This well-known algorithm solves a variety of problems because it is simple and straightforward. The main concept is partitioning the objects into disjointed clusters by minimizing the value of the similarity between the target object and the center of the clusters.

The step of k-means clustering is:

a) Choose randomly k items as initial centers.

b) Distance computation between the rest of items and k centers. Assemble items into the closest cluster centers.

c) The mean is computed for all the items within one cluster. The new center is the average of each cluster.

d) The difference between the newly calculated center and the pervious center in the same cluster is computed. If the threshold is greater than the difference value then repetition must continue and the new centers is replaced by the previous one (return to b) and continue). If maximum number of repetitions has reached, the algorithm is terminated [2].

5. The Proposed Recommendation System

5.1 Dataset Description

A MovieLens 100k dataset is used from the GroupLens Research Group web-based research recommender system.it consists of 100,000 ratings from 943 users on 1682 movies. In this dataset at least 20 movies each user rated. The fields of MovieLens data set are (as shown in Fig.(1)): the user ID, Movie ID, rating and Timestamp field that represent time in seconds. This dataset was chosen because it has different scaling values for users and different number of ratings for each user. The file ratings.dat is conducted for information extraction.

user ID	item ID	Rating	Timestamp
1	17	3	875073198
1	47	4	875072125
1	64	5	875072404
1	90	4	878542300
1	92	3	876892425
1	113	5	878542738
1	222	4	878873388
1	227	4	876892946
1	228	5	878543541
1	253	5	874965970
2	257	4	888551062
2	270	4	888551745
2	299	4	888550774
2	301	4	888550631
2	303	4	888550774
2	307	3	888550066
2	308	3	888979945
2	313	5	888552084

Fig.(1): the MovieLens Svc file Snapshot.

5.2 Model Structure

The proposed recommendation model consists of three phases as shown in Fig.(2):

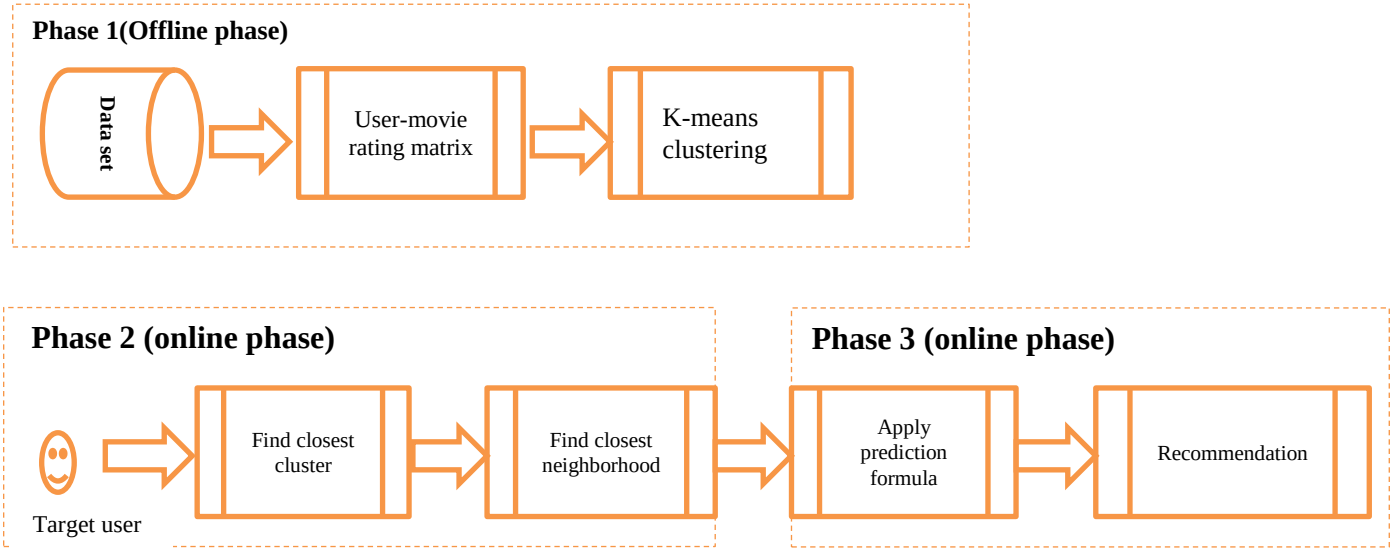


Fig.(2): The proposed cluster-based model.

5.2.1 The Preprocessing Phase

This phase is done offline. The user-movie rating matrix is constructed using the MovieLens dataset. After matrix formulation, K-means clustering is applied on the matrix to build a cluster-based model. User Clustering results in creating sub-matrices from the entire user-movie rating matrix. The essential idea of K-means clustering is to arrange a given set of MovieLens users into K number of disjoint clusters. The value of K is given a priori. The distinct steps of cluster-based model are:

1. K user vectors are chosen randomly as centroids from the user-movie rating matrix, one for each cluster.
2. The rest of user vectors are taken and a mapping function (Euclidean distance) is used to map them to the closest centroids.
3. This loop will continue until all the users are involved in the designated clusters. Then a recalculation of a new centroid vectors is done. As a result of the loop, K centroid vectors may change their location in step by step routine. Finally, a state will reach when the centroid vectors do not change.

The Clusters will be used as a neighborhood for the target user as it will be presented in next section.

5.2.2 The Similarity Phase

A target user enters as input to the proposed cluster-based model to get recommendations online. A similarity calculation is done between the target user (a) with each cluster centroid using the similarity formula (Eq.3) below to find the closest cluster to the target user.

$$Sim(U_a, C_u^j) = \frac{\sum_{i \in I} (r_{u_a,i} - \overline{r_{u_a}})(r_{C,i} - \overline{r_C})}{\sqrt{\sum_{i \in I} (r_{u_a,i} - \overline{r_{u_a}})2} \sqrt{\sum_{i \in I} (r_{C,i} - \overline{r_C})2}} \quad \text{.....(Eq. 3)}$$

Where i is item belongs to I, C_u^j means the j^{th} user cluster.

After selecting the closest cluster, calculate similarity using Pearson correlation coefficient similarity measure (Eq.1) as presented in section 3.3.1 between the target user (a) and users belonging to the closest cluster.

The 10 highest similarities are chosen as the target user's neighbors. Fig.(3) below shows the implementation of user-user similarity matrix using Pearson correlation coefficient.

	1	2	3	4	5	6	7	8	9	10
1	1	0.0919	-0.0052	0.0200	0.2922	0.3306	0.3235	0.2568	0.0565	0.2776
2	0.0919	1	0.1031	0.1046	0.0018	0.1803	0.0279	0.0603	0.0837	0.0739
3	-0.0052	0.1031	1	0.2269	-0.0199	0.0251	-0.0082	0.0504	-0.0124	0.0064
4	0.0200	0.1046	0.2269	1.0000	-0.0125	-0.0312	0.0110	0.1429	-0.0073	-0.0116
5	0.2922	0.0018	-0.0199	-0.0125	1.0000	0.1540	0.2710	0.1910	0.0593	0.1089
6	0.3306	0.1803	0.0251	-0.0312	0.1540	1.0000	0.3808	0.1072	0.0847	0.4610
7	0.3235	0.0279	-0.0082	0.0110	0.2710	0.3808	1.0000	0.2098	0.0881	0.3730
8	0.2568	0.0603	0.0504	0.1429	0.1910	0.1072	0.2098	1.0000	0.0156	0.1562
9	0.0565	0.0837	-0.0124	-0.0073	0.0593	0.0847	0.0881	0.0156	1	0.1428
10	0.2776	0.0739	0.0064	-0.0116	0.1089	0.4610	0.3730	0.1562	0.1428	1.0000
11	0.2237	0.0683	0.0354	0.0696	0.2584	0.1594	0.2782	0.0998	0.0039	0.1310
12	0.2417	0.0786	0.0115	0.0121	0.0495	0.1718	0.1848	0.1094	0.1705	0.1933
13	0.2717	0.1217	0.1192	0.0318	0.2229	0.3304	0.3842	0.2216	0.1016	0.3588
14	0.2481	0.1649	0.0160	0.0031	0.1913	0.2911	0.2028	0.1265	0.1152	0.2276
15	0.0961	0.3682	0.0699	0.1057	0.0127	0.1144	0.0084	0.0293	0.0544	0.0371
16	0.2921	0.1006	0.0186	0.0305	0.2009	0.2925	0.3449	0.2011	0.0613	0.3933
17	0.1402	0.1860	-0.0146	-0.0086	0.0298	0.1376	0.0376	0.0568	0.2178	0.0941
18	0.3728	0.1213	-0.0228	-0.0223	0.1860	0.4906	0.3983	0.0929	0.0692	0.4032
19	0.0284	0.0245	0.0893	0.0398	0.0322	0.0826	0.0681	0.0247	0.0634	0.0411

Fig.(3): user-user similarity matrix.

5.2.3 The Prediction Phase

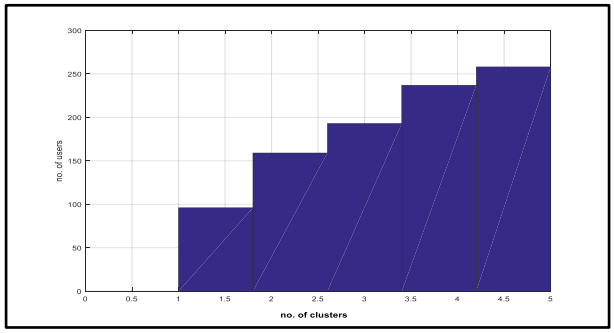
After a similarity computation phase, a prediction is generated by aggregating weighted average of deviations from the neighborhood users within the same cluster (that have highest similarities). The formula (Eq.2) presented in section 3.3.2 is used for predicting rating for the target user. Then movies with a high rating prediction are chosen as recommended movies.

6. Results and Discussion

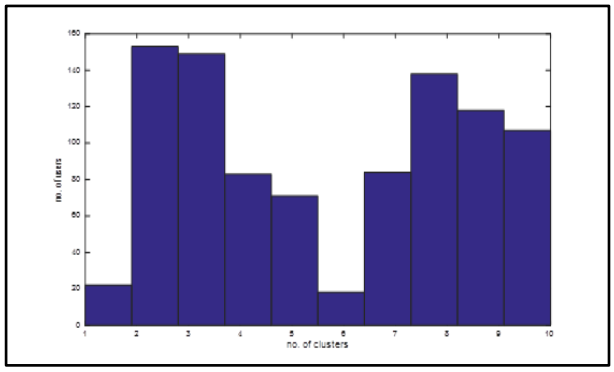
The proposed model is tested by conducting experiments on the most popular MovieLens dataset. Experiment is performed using MATLAB tool. In MovieLens Dataset rating data files have at least three columns: the user ID, the item ID, and the rating value. Users are clustered and distributed using K-means clustering method. The results shown that using K=5 and Pearson correlation Distance measure instead of Euclidian distance lead to uniform distribution.

In which cluster1 has 96 users, cluster 2 has 159 users, cluster 3 has 193 users, cluster 4 has 237 users and cluster 5 has 258 users. So no exhaustive time needed for similarity calculation when finding similar users within the cluster. When K=10, the number of users are not distributed evenly as shown in Fig.(4).

Table (1) shows the results after implementing the prediction formula for users chosen randomly to predict their ratings to movies they did rate and movies they did not rate. The results show for users belonging to different clusters. Using Pearson correlation, the predicted ratings are relatively similar to the real ratings.



(a)



(b)

Fig.(4)(a): Users distributed on 5 clusters (b) users distributed on 10clusters.

Table (1)
Prediction results.

users	User 1 prediction to:				User 2 prediction to:		User4 prediction to:	User 5 Prediction to:	
movies	Movie ID.2	Movie ID.3	Movie ID.4	Movie ID.5	Movie ID.1	Movie ID.10	Movie ID.11	Movie ID.42	Movie ID.63
Prediction rating	3.03≅3	3.53≅4	3.64≅4	3.30≅3	4.09≅4	3.84≅4	4.49≅5	3.12≅3	2.52≅3
Real rating	3	4	3	3	4	2	0	5	0

7-Conclusion

In this paper, K-means clustering method is explored to address the scalability issue which is a fundamental challenge in recommender systems. Applying K-means clustering offline on user-movie rating matrix reduced the sparseness and reduced the scalability problem of the model since the computation of finding similar users to the target user is only calculated for users within same cluster, thus reducing the target user's neighbor number. If a movie is selected by

these users, it will be suitable to the target user. According to the prediction results, Pearson correlation coefficient similarity measure achieved relatively good results to find the closest neighbors to the target user.

References

[1] Khurana P., Parveen S., "Effective Hybrid Recommender Approach using Improved K-means And Similarity", international journal of computer trends and technology (IJCTT), 36(3), 2016.

[2] Liao Q., Yang F. and Zhao J., "An improved parallel K-means clustering algorithm with MapReduce", IEEE, In Communication Technology (ICCT), 15th IEEE International Conference (764-768), 2013, November.

[3] Kumar M., Yadav D.K., Singh A. and Gupta V.K., "A movie recommender system: Movrec", International Journal of Computer Applications, 124(3), 2015.

[4] Zebin Wu, Yan Chen and Taoying Li, "Personalized Recommendation Based On The Improved Similarity and Fuzzy Clustering", IEEE, The National Natural Science Foundation of China (No.71271034), 2015.

[5] Yao G., Cai L., "User-Based and Item-Based Collaborative Filtering Recommendation Algorithms Design", University of California, San Diego, 2017.

[6] Al-Bakri N.F., Hashim S.H., "Reducing Data Sparsity in Recommender Systems", Journal of Al-Nahrain University-Science, 21(2), (pp.138-147), 2018.

[7] Bandyopadhyay S., Saha S., "Unsupervised classification: similarity measures, classical and metaheuristic approaches, and applications", Springer Science & Business Media, 2012.