Journal of Engineering and Applied Sciences 13 (Special Issue 5): 4637-4643, 2018
ISSN: 1816-949X

Demographic Features Cooperationfor Enhancing Collaborative Filtering Recommender System

Zainab Khairallah Kadhim and Huda Naji Nawaf
Department of Network, Faculty of Information Technology,
University of Babylon, Hillah, Iraq

Abstract: Collaborative filtering is one of prevalent successful methods of recommender system. In this study, a prediction model for homophily clustering of users has been built to improve the collaborative filtering recommender system. The general framework mainly consists of two phases: Firstly, detect communities in homophily networks by using Partitioning Around Medoids (PAM) clustering algorithm. Secondly, building naive bayes model by calculating the conditional probability foruser's demography attributes. The experiments have been applied on two real world datasets 100 K and 1 M that published by grouplens. Finally, precision, recall and F-measure metrics have been used to evaluate the top-N recommendation lists. The empirical results can provide a recommendation in a best manner also, the results have compared with other research study.

Key words: Recommender system, collaborative filtering, homophily, demographic, gender parameter, K-medoids age parameter

INTRODUCTION

The WWW (World Wide Web) has become the mainsource for many users in the world. Also, the various data resources and different processing operations make the tremendous amount of the information on a web that causes the difficulty of a person for a researching all the productions in the specific website and making a decision to select one of them. At the same time, the development of technologies and information overload has been origin concern among interest users and content providers. The user finding has difficulty in new interesting things and the provider looked for a tool to increase the customer's trust, loyalty, increase sales and obtain more knowledge about customers (Belkin and Croft, 1992).

The next generation of development in the information filtering that represented in the recommender system. Recommender system is a technique to solve the problem of the information overload and help the users to make decisions when they are many choices to select based on the user's preferences, service or observed behavior about an item (Konstan and Riedl, 2012). One ofthe personalized systemsis a recommender system that offers different items for each customer in the online or offline store. This difference of the selected items basedon the preferences of the past history of the user or other similar users. The system uses the various related information of users to recommend the items according to the type of approaches of the recommender systems that are used (Jannach *et al*., 2011).

Collaborative Filtering (CF) has become one of the most used approaches of recommender system to providing personalized services for users. This methodbased on the rating of the other similar users to items and this information used to predicate items to the active user. Collaborative filtering can often be grouped as being either: memory-based or model-based (Breese *et al*., 1998). Memory-based used on the entire dataset on contrary the model-based build a model and using the data mining and machine learning to extract information from dataset. In recent years, recommender system has become significantly populardue to its importance in the many fields, entertainment (recommended movie, music, games and etc.) social network recommended (feeds facebook, Twitter tweets, suggest friendsand, etc.) e-Commerce (recommended to buy products) and many other fields (Ricci *et al*., 2015).

Homophily is a common concept in the social network. It means to associate the similar people together and make relationships according to share commoninterpersonal characteristics and connect the

Corresponding Author: Zainab Khairallah Kadhim, Department of Network, Faculty of Information Technology, University of Babylon, Hillah, Iraq

similar people at a higher rate than among dissimilar people (Mcpherson *et al*., 2001). While the demographic recommender system includes (age, gender, nationality, location, education and other properties) and uses the properties of a user to produce the list of recommendations. Worthy to mention that some studies based on the demographic correlation between users lead to enhancing the quality of prediction recommendation and explained the important effect of using demographic information on the recommender system (Pazzani, 1999; Chen and He, 2009).

Literature review: Many of the methods proposed to improve the collaborative filtering system, the core of memory-based collaborative filtering is a similarity measure, so, Cheng *et al*. (2015) triedthe enhancement the CF by building two measures. The first model based on the different movie genres that the user has viewed while the other similarity based on the average ratings to improve the recommender system (Cheng *et al*., 2015). Zhang *et al*. (2014) suggested a new similarity measure based on random walk with choice where this measure takes into consideration the length of vectors that the traditional measures take into account only the direction of the vectors. Computing this similarity can solve sparsity problems which can develop the prediction of collaborative filtering (Zhang *et al*., 2014). While Das *et al*. (2014) attempted to handle the scalability problem of recommender system and reducing the processing time. Their proposed methodbased on DBSCAN clustering algorithmto partition the users into groups (Das *et al*., 2014). The other researchers Ju and Xu. (2013) improved the collaborative filtering based on Artificial Bee Colony (ABC) algorithm to outcome the local optimal problem of k-mean algorithm by selecting the optimal center of the initial centers of the k-mean algorithm. Also, they modified cosine similarityby computing the common between the users in the same cluster (Ju and Xu, 2013). Moreover, some researchers have suggested new method to improve recommendation by taking into count the changes of the behavior of the users over the time.

Also, Wang and Tan (2011) proposed a new collaborative filtering approach based on improving the original Naive Bayes classifier and applied k-NN Nearest Neighbor algorithm. Generally, the proposed approach is an appropriate when the length of recommendation list is long (Wang and Tan, 2011). Beel *et al*. (2013) manifested the importance and effect of the demographic information and the characteristics of the personal users in improving the prediction of the recommendations and an evaluation ofthe recommender system. Age information has influenced strongly on Click Through Rates (CTR) measure. But the gender had only a marginal effect (Beel *et al*., 2013). While Aygun and Okyay (2015) noticed the idea of a gap generation between the age of the users and hence, they suggested the similarity measure based on the age parameter and the traditional pearson similarity to improve the collaborative filtering. The comparison between the traditional similarity measures and the suggested showed the improvement of the age parameter. But this method is a memory-based method (Aygun and Okyay, 2015).

Another researchers, Swamy and Reddy (2015) tried to increase the diversity of items and the accuracy of the collaborative filtering. They suggested association rules recommender system focused on the distribution of the items. The items were classified as a concept hierarchy. By using the class of item can be selected the diversified rank. They also used the model-based and top-N recommendation to suggest recommendations to users (Swamy and Reddy, 2015). Finally, You *et al*. (2015) increasethe quality of prediction CF and to solve the problems of (cold start, sparsity and scalability). They proposeda method based onpartition of the items into several clusters and they applied one weighted slope on each cluster to predict the rating ofitems. DBSCAN algorithm is used to group the items. The top rating of items can be the suggested to active users. This method predicts the values of rating for the unknown items (You *et al*., 2015).

MATERIALS AND METHODS

The proposed model presents a recommendation system that recommends items to users based on demographic information (age and gender) and homophily network. Worth to mention that our research by (Khairallah and Vawaf, 2016) has been proposed to incorporate the demographic information using only age and homophily network. However, the age and gender have been presented in this research to fully aware which parameter has more effect on improving the recommendation system. The model consists of four main stages: pre-processing, community detection, Prediction model, evaluation predictionas shownina (Fig. 1).

The first stage in the proposed method is the pre-processing. The pre-processing stage has extracted a sequence of items (movies) have been viewed from the original movielens dataset for each user. The data consist

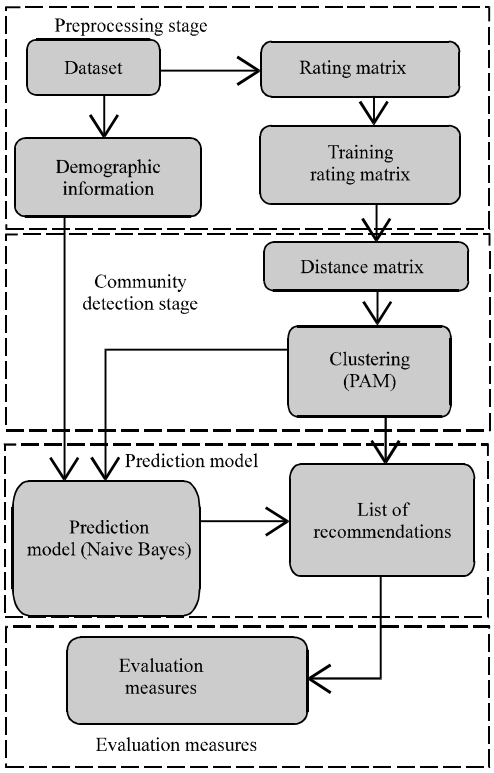

Fig. 1: The block diagram of the proposed method

of the four features (Users U, Items It, Rating R, time). The data divided into training and testing data based on user sequence ratings. Additionally, rating matrix (user-item matrix) has been created within pre-processing stage for the two training and testing data the empty entry in the rating matrix represents the sparse data which means the user did not watch this item or the user watched it but he/she never evaluated it. In the two datasets which have been used to evaluate the proposed method, the number of ratings by each user must be equal to or more than 20 items.

Community detection is the second stage in proposed methodology which includes the following two steps: computedthe distance measure on the rating matrix and applied the PAM (Patriating Around Medoids) algorithm clustering. PAM (Partitioning Around Medoids) is a partitioning clustering algorithm that based on the location of the center as a algorithm 1. PAM is more robust than a k-mean algorithm in isolating the outliers and it does not depend onthe order of objects. Some approaches have been suggested to avoid the number of a clusters issue such as; Silhouettes plot which has been applied in this research. The Silhouettes coefficient is a geographic representation that can determine the optimal number of clusters. It can prevent us from drawing the

Table 1: Categories of ages

Categories	Ranges of age
1	Under 18
2	18-23
3	24-28
4	29-33
5	34-38
6	39-43
7	44-48
8	49-53
9	54-58
10	59-63
11	64+

wrong conclusions for the selection of the number of clusters (Kaufman and Rousseeuw, 1990). The third stage is the prediction model which is the important stage in building the model. The building of the model need to use the Naive Bayes Model which has been applied to predict the recommendation lists with various lengths. The assumption that all variables X_i are mutually independent given a "special" variable C. The joint distributionis then given compactly by Lowd and Domingos (2005).

Algorithm 1; PAM algorithm:

Kmedoids (PAM Partitioning Around Medoids)
Input:
kthe number of clusters, Da set of n objects
Output:
a set of k clusters
Begin
Select k initial medoids arbitrarily from D. Repeat until no changes in clusters medoids
For each object
Calculate its distance from the medoid of each cluster. Assign the object to the cluster with nearest medoid
For each medoidm do
Min equal sum of the distances from m to the other objects
For each non-medoid objec to do
Calculate the sum of the distances from o to the other objects
If the sum is less than min
Then swap m and o and update min
End

The probability of $(X_1, X_2, ..., X_n)$ given as : the model uses the age of user (X_1) , gender of user (X_2) and the last items that have been viewed by a user (X_3) as the prior knowledge to predict items (C). The co-occurrence matrix has been produced for each cluster as the main step of the model. In each cluster, the items have been viewed by users of this cluster should produce the items of the co-occurrence matrix. The three co-occurrence matrix associate between the new items (the item that has been never rated by the active user) and the last item of the user and the last two matrices associate between the new items, age and gender of users. Generally, the last item and demographic information (age and gender) are considered as probability conditions in the NB prediction model. The ages have been divided into 11 ranges as a (Table 1).

Table 2: Confusion matrix

Predictive model	
Yes	No.
Actual recommeded No.	
Yes True Positive (TP)	False Negative (FN)
False Positive (FP)	True Negative (TN)

Table 3: The information about the datasets

Data set	Users	Items	Rating	Values of rating
100 K Moivelens 1	943	1682	100000	1-5
1M Moivelens 2	6040	3706	1000209	1-5

The output of occurrence probabilities for each item has been createdfor each user in each cluster. Subsequently, the probabilities have been sorted decently.

Demographic information and homophily network are the core of the proposed method. Both age and gender parameters have been taken into account in building the Naive Bayes Model (Khairallah and Nawaf, 2016). In fact, Naive Bayes Model has been built using each parameter alone then associate them in one model. In this stage, the proposed model generates list of different recommendations for each user. In the final stage, the accuracy of the model has been evaluated by using the popular measures such as (precision, recall, F-measure) and accuracy. The rating matrix in the pre-processing step is divided into training and testing data. The dividing of the training and the testingdata is 70 and 30% of the user sequence, respectively where theitems that have been rated recently by each user have been assigned as a test set. In other words, these items can be considered as a future behaviour of users. It hence, the confusion table has been applied to acquire the above measures in a (Table 2).

The experimental results: The experiments performed on two movielens datasets with different sizes. Movielens dataset was gathered by the Grouplens research project at the University of Minnesota. The proposed system has been applied on two data sets, the dataset is divided into training and test sets with percentage 70 and 30%, where the detail information about these datasets have been stated in a (Table 3). In the two datasets each user has at leasta 20 rating.

The classification measures (Precision, recall and F-measure) based on the confusion matrix have be used to evaluate the recommender system. The values of these measures cannot be interpreted as absolute measures, then these values can be compared with various prediction algorithms on the same dataset (Cremonesi *et al*., 2008).

Two experiments have been performed to show the superiority of the proposed system over baseline (The fixed experiment without clustering the users and without using demographic information). The first experiment showed the effect of the using combination the ageand gender properties. It clusters the users into different size of group and applied the prediction model (Naive Bayes) with two demographic information on each cluster as a (Fig. 2).

In the second experiment as a (Fig. 3), the prediction modelhas be applied on clusterswhich include the users who are similar in their interest (homophily). Indeed these clusters have been obtained using PAM. It is similar with the first experiment but it not contains the main factor (demographic information) to explain the important of using the age and gender features to improving the collaborative filtering.

Table 4 and 5 show the percentages of items with highest ratings 4 and 5 in each list that recommended by the proposed method with different lengths. It may be noted that the recommendation list should include the highest rating items in test set which reflect the accuracy of the proposed system, thus, it meets the satisfaction of users in real world.

Additionally, the performance of the proposed method is superior over the existing works. Important to say, the comparison with the other researches has been

Table 4: The percentages of 4 and 5 ratings for different recommendation list length

Dataset	100 K Movielens dataset (%)	1M Movielens dataset (%)
1	69.2	76.6
5	67.8	75.8
10	67.2	75.2
15	66.5	75.0
20	66.2	74.9
25	66.2	74.9
30	66.2	74.9
35	66.2	74.9
40	66.2	74.9

Table 5: Compare our proposed method with other methods (a multi-level CF (Platidis *et al*., 2016), S1 model by Cheng *et al*. (2015) on the first dataset 100 K Movielens

Lists	The method	Precision	Recall	F-measure
5	Our method	**0.457**	**0.075**	**0.112**
	A multi-level CF Polatidis and Georgia (2016)	0.050	0.060	0.055
10	Our method	**0.411**	**0.129**	**0.168**
	A multi-level CF Polatidis and Georgia (2016) S1 Model	0.060	0.070	0.065
	by Cheng *et al*. (2015)	0.090	0.030	0.035
15	Our method S1 Model	**0.380**	**0.173**	**0.194**
	by Cheng *et al*. (2015)	0.080	0.038	0.040
20	Our method S1 Model	**0.355**	**0.213**	**0.216**
	by Cheng *et al*. (2015)	0.075	0.048	0.045
25	Our method S1 Model	**0.337**	**0.250**	**0.233**
	by Cheng *et al*. (2015)	0.072	0.058	0.050
30	Our method S1 Model	**0.321**	**0.283**	**0.244**
	by Cheng *et al*. (2015)	0.070	0.068	0.052
35	Our method S1 Model	**0.309**	**0.316**	**0.254**
	by Cheng *et al*. (2015)	0.068	0.075	0.056

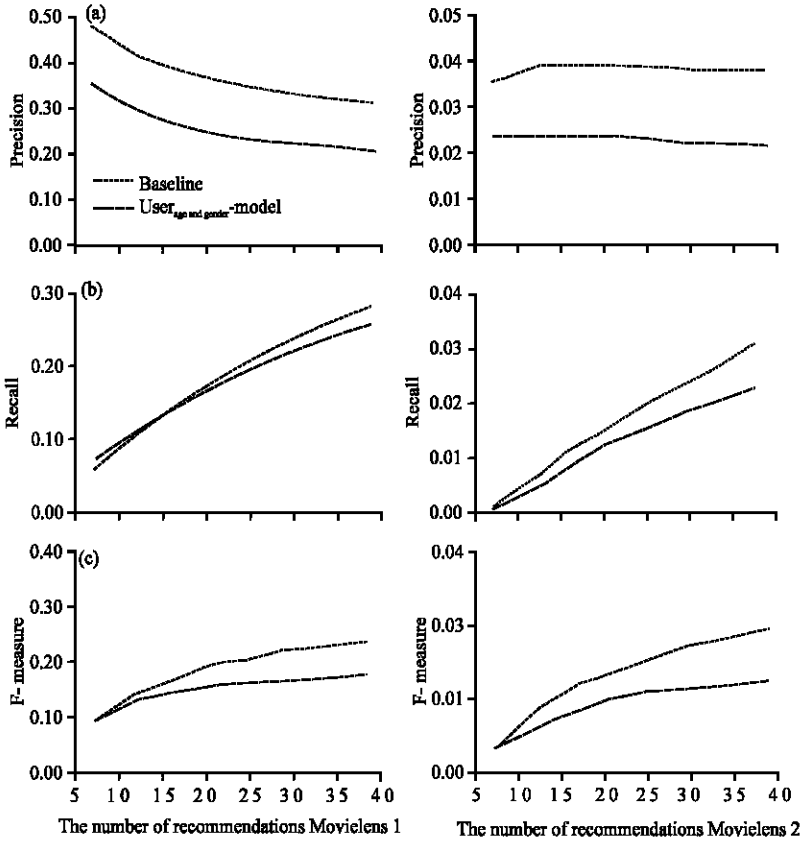

Fig. 2: The performance of the proposed method against the baseline

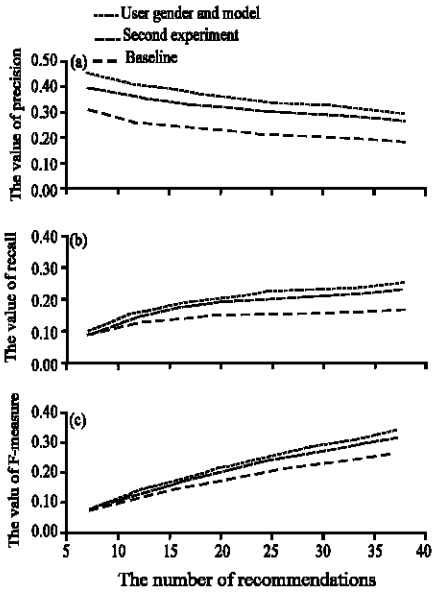

Fig. 3: The comparison among the proposed method and the baseline and second experiment on the100 K Movielens dataset: a) Precision measure; b) Recall measure and c) F-measure

constrained according to the published results in their works. The proposed method has compared based on the length of the TopN recommendation lists as it is shown in Table 5.

RESULTS AND DISCUSSION

The performance of the proposed method (green curve) is best as shown in Fig. 2 which means when the age and gender have been combined in one model the performancehas been improved.

The first experiment use age and gender attribute together (green curve) to showthe increasing of mapping the propose items (suggested items) to users with the their test items make the develop the performance of precision measure. After that, improve the F-measure and the collaborative filtering system. The second experiment explains the good effect of finding demographic information within the preferences of users. It represents (blue curve) and compares with two experiments the baseline (red curve) and $User_{gend\ and\ age}$ model (green curve).

The demographic information show the positive effect on enhancing the collaborative filtering in all classification measure (precision, recall and F-measure). As for the (Table 4), the percentages illustrate that recommendation lists with different lengths comprise the highest percent of important items have 4 and 5 ratings in both datasets which confirm on the accuracy of the proposed method. The highest value of rating is equal 4 from other value on the first dataset and 66% from rating items is about (4 and 5) rating for user.

Also the highest rating equal 5 in the second dataset and 75% from rating was about (4, 5) rating as a (Table 4). The proposed method was very nearly from the preferences and interesting items of users.

CONCLUSION

We present two factors that can be affected on the prediction of recommendation system, the cluster model and demographic information. In this study, a prediction model for homophily clustering of users has been built to improve the collaborative filtering recommender system. In addition, take the demographic factor affect positively on the accuracy of the system and the personal information is considered as a good factor in determining the preferences. In concluding, It has been found that the age and gender together factor in homophily communities improves the performance of the system.

Finally we plan to use other demographic information such as nationality, profession, field of the research, location and other properties to improve the collaborative filtering.

REFERENCES

Aygun, S. and S. Okyay, 2015. Improving the pearson similarity equation for recommender systems by age parameter. Proceedings of the IEEE 3rd Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE), November 13-14, 2015, IEEE, Riga, Latvia, ISBN:978-1-5090-1202-2, pp: 1-6.

Beel, J., S. Langer, A. Nurnberger and M. Genzmehr, 2013. The impact of demographics (age and gender) and other user-characteristics on evaluating recommender systems. Proceedings of the International Conference on Theory and Practice of Digital Libraries, September 22-26, 2013, Springer, Berlin, Germany, ISBN:978-3-642-40500-6, pp: 396-400.

Belkin, N.J. and W.B. Croft, 1992. Information filtering and information retrieval: Two sides of the same coin?. Commun. ACM, 35: 29-38.

Breese, J.S., D. Heckerman and C. Kadie, 1998. Empirical analysis of predictive algorithms for collaborative filtering. Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, Jul 24-26, 1998, Madison, WI., pp: 43-52.

Chen, T. and L. He, 2009. Collaborative filtering based on demographic attribute vector. Proceedings of the International Conference on Future Computer and Communication FCC'09, June 6-7, 2009, IEEE, Wuhan, China, ISBN:978-0-7695-3676-7, pp: 225-229.

Cheng, Q., X. Wang, D. Yin, Y. Niu and X. Xiang et al., 2015. The new similarity measure based on user preference models for collaborative filtering. Proceedings of the IEEE International Conference on Information and Automation, August 8-10, 2015, IEEE, New York, USA., ISBN:978-1-4673-9104-7, pp: 577-582.

Cremonesi, P., R. Turrin, E. Lentini and M. Matteucci, 2008. An evaluation methodology for collaborative recommender systems. Proceedings of the International Conference on Automated solutions for Cross Media Content and Multi-channel Distribution, November 17-19, 2008, IEEE, New York, USA., ISBN:978-0-7695-3406-0, pp: 224-231.

Das, J., P. Mukherjee, S. Majumder and P. Gupta, 2014. Clustering-based recommender system using principles of voting theory. Proceedings of the International Conference on Contemporary Computing and Informatics (IC3I), November 27-29, 2014, IEEE, Mysore, India, ISBN:978-1-4799-6630-1, pp: 230-235.

Jannach, D., M. Zanker, A. Felfernig and G. Friedrich, 2011. Recommender System: An Introduction. Cambridge University Press, New York, USA.,.

Ju, C. and C. Xu, 2013. A new collaborative recommendation approach based on users clustering using artificial bee colony algorithm. Sci. World J., 2013: 1-9.

Kaufman, L. and P. Rousseeuw, 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, Hoboken, New Jersey, Pages: 342.

Khairallah, Z. and H.N. Nawaf, 2016. Improving recommendation system based on homophily principle and demographic. Res. J. Appl. Sci., 11: 1102-1106.

Konstan, J.A. and J. Riedl, 2012. Recommender systems: From algorithms to user experience. User Model. User Adapted Interact., 22: 101-123.

Lowd, D. and P. Domingos, 2005. Naive bayes models for probability estimation. Proceedings of the 22nd International Conference on Machine learning, August 07-11, 2005, ACM, Bonn, Germany, ISBN:1-59593-180-5, pp: 529-536.

McPherson, M., L.S. Lovin and J.M. Cook, 2001. Birds of a feather: Homophily in social networks. Annu. Rev. Sociology, 27: 415-444.

Pazzani, M., 1999. A framework for collaborative filtering, content-based and demographic filtering. Artif. Intell. Rev., 13: 393-408.

Polatidis, N. and C.K. Georgiadis, 2016. A multi-level collaborative filtering method that improves recommendations. Expert Syst. Appl., 48: 100-110.

Ricci, F., L. Rokach and B. Shapira, 2015. Recommender System Handbook. 2nd Edn., Springer, Heidelberg, Germany, ISBN:978-1-4899-7636-9, Pages: 1003.

Swamy, M.K. and P.K. Reddy, 2015. Improving diversity performance of association rule based recommender systems. Proceedings of the International Conference on Database and Expert Systems Applications, September 1-4, 2015, Springer, Cham, ISBN:978-3-319-22848-8, pp: 499-508.

Wang, K. and Y. Tan, 2011. A new collaborative filtering recommendation approach based on naive bayesian method. Proceedings of the International Conference on Advances in Swarm Intelligence, June 12-15, 2011, Springer, Chongqing, China, ISBN:978-3-642-21523-0, pp: 218-227.

You, H., H. Li, Y. Wang and Q. Zhao, 2015. An improved collaborative filtering recommendation algorithm combining item clustering and slope one scheme. Proceedings of the International Multi Conference on Engineers and Computer Scientists Vol. 1, March 18-20, 2015, IAENG Publisher, Hong Kong, ISBN:978-988-19253-2-9, pp: 1473-1476.

Zhang, X., C. Mi, X. Shan and J. Ma, 2014. Collaborative filtering algorithm based on random walk with choice. Proceedings of the 2nd International Conference on Software Engineering, Knowledge Engineering and Information Engineering (SEKEIE 2014), August 5-6, 2014, Atlantis Press, Singapore, pp: 192-196.