Research Journal of Applied Sciences 11 (10): 1102-1106, 2016
ISSN: 1815-932X

Improving Recommendation System Based on Homophily Principle and Demographic

Zainab Khairallah and Huda Naji Nawaf
Department of Software, Babylon University, Babil, Iraq

Abstract: Collaborative filtering is one of the prevalent successful approaches in the Recommender systems to predicate items to users based on rating matrix and mitigate the difficulty of finding interesting things on the spider's web. In this paper, we percent a Naïve Bayes model by taking into account the similarity in preferences (homophily) among the users and attributes of users (demographic) as a prior knowledge to enhance the prediction accuracy of collaborative filtering. Experiments are implemented on Movielens datasets 100K and 1M. The results show that the system can provide a recommendation in a best manner.

Key words: Collaborative filtering, homophily, demographic, Naive Bayes classifier, clustering, K-medoids

INTRODUCTION

The amount of information on the web has long been growing very rapidly more than our ability to analyze or organize it. Moreover, the development of technologies and information overload has been origin concern among interest users and content providers. The difficulty of a user to find new interesting things and the provider looked for a tool to increase trust and customer loyalty, increase sales and obtain more knowledge about customers. Recommender system is a powerful tool that can help users to navigate through the massive information of items or products. The recommender system is applied in the e-commerce (Amazon, Netflix and etc.) to suggest a list of products (books, Movies, CDs and etc.) to customers or to find interesting items or friends such as in the social network field and it can apply in many fields (Sun *et al*., 2012).

Recommender system is a branch of information filtering that it uses several of machine learning and statistical approaches to predicate items to users based on certain methods (Ricci *et al*., 2015). A large number of items are one of the challenges in the recommender system, so the use of the cluster technique to reduce the proposed items to a user based on the domain of items of each cluster. The information about collaborative filtering that based on the preferences of users can be improved. The collaborative filtering (Hofmann and Puzicha, 1999) based on the rating the user to items and it can solve many limitations of content-based. One of this limitation is the difficulty of availability of properties of items. Furthermore, over-specification of items. Collaborative filtering can often be grouped as being either: Memory-based or Model-based (Breese *et al*., 1998). The main key of memory-based (nearest- neighbor) has found similarity between users and applied prediction function to predicate items to a user. Model-based extracts the information from the dataset and builds a model by using the machine and data learning techniques such as clustering or probabilistic and other approaches (Ricci *et al*., 2015). In this study, weexploit the demographic information of users in addition to homophily network to improve the performance of recommendation system.

Literature review: There are many researchers attempted to improve Collaborative filtering by proposing new similarity measures that used Entropy-based, Fuzzy-weighted (Shamri *et al*., 2014) and other methods (Huang *et al*., 2015; Cheng *et al*., 2015). Other suggestions to resolve the cold start problem to enhance the prediction of recommendation system (Ahn, 2008; Gogna and Majumdar, 2015). Another multi-level collaborative filtering method to improve the recommendation (Huang *et al*., 2015).

Clustering algorithm: One of the challenges of k-nearest neighbor is sensitive to sparse that leads to improve other so the clustering techniques to solve this issue.Clustering isan unsupervised learning that uses the similarity or distance measu res to capture all the users into limited and discrete sets (Treerattanapitak and Jaruskulchai, 2012). However, several existing works in the collaborative filtering field use the k-mean algorithm. Extended k-mean

Corresponding Author: Zainab Khairallah, Department of Information Networks, Babylon University, Babil, Iraq

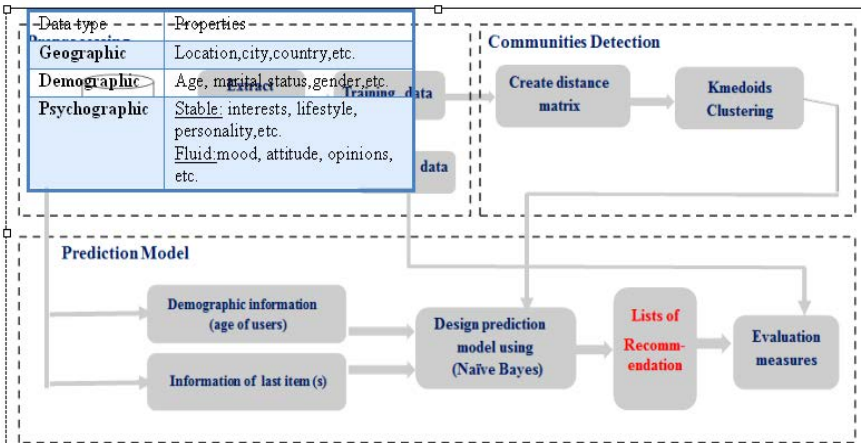

Fig. 1: User profile classification

as been proposed by Tang (Wang and Tan, 2011) for build the item's precise category system while in (Treerattanapitak and Jaruskulchai, 2012), the exponential fuzzy c-mean used.

PAM (Partitioning Around Medoids) is a partitioning clustering algorithm that based on the location of the center (Kaufman and Rousseeuw, 1990). It is more robust than a k-mean algorithm to isolate the outliers and it doesn't depend on the order of objects. The key issue in some clustering methods is how determines the number of a clusters. Some approaches have been suggested to avoid this issue such as; silhouettes plot which has been applied in this work. Worthy to mention, the drawback of PAM high time complexity (Kaufman and Rousseeuw, 1990). PAM consists of main two phases: Build and swap phases and it has shown in Fig. 1 (Pillay *et al*., 2015).

Bayesian Classifier: It is a statistical approach that's based on probability conditional used to learn the model to classify the data into two classes for instance;like or don't like. The most common type of Bayesian classifier is Naïve Bayes (Ricci *et al*., 2015). Naïve Bayes assumes theclass label and attributes as a random variable based on the Bayes Theorem. For instance the event (c) conditional probability of many events (d_1, d_2,d_3,…,d_n) = P(c/d1, d2,d3,....,dn) and using the statistical Bayes Theorem. Generally, the Naïve Bayes Eq. 1 (Murty and Devi, 2011).

$$p(c / d_j)P(d_j) = P(d_j / c) = p(c) \quad (1)$$

$P(d_j \mid c)$ = Probability of instance c being in class d_j
$P(c \mid d_j)$ = Probability of generating instance c has given class d_j
$P(d_j)$ = Probability of occurrence of class d_j
$P(c)$ = Probability of instance d is occurring

This classifier has isolated the noise data and the ability to delete the instance through the probability calculation which leads to solving the missing value problem. TheNaïve Bayes model has been proposed by several researchers existing researches in this field (Pazzani, 1999; Wang and Tan, 2011; Miyahara and Pazzani, 2000a, b).

User profile modeling: Clema suggested that the user profile can be classified into three as states in Fig.2. The demographic recommender system uses the properties of a user to produce the list of recommendations.

The resercher (s) Beel *et al*. (2013) have explained the important effect of demographic information (age property) on click-through rate when they utilized on the recommender system.

MATERIALS AND METHODS

The proposed model: The proposed model presents a movie recommendation system that recommends movies to users based on their personal information and their rating of the other items. In proposing method the shared interests among users (homophily network). The model consists of three main parts: pre-processing, communities detection and prediction model as shown in Fig. 3.

Preprocessing: The first part of this model is preprocessing to prepare the data for processing. In particular, the input data consists of the users Ui, the items Ij and the Rratings (R) which have been given by the users for items. This phase includes creating the

sequence for each user represents the movies have been viewed by them, then dividing the data into the training and testing data. The items of training data have been chronologically arranged. The rating matrix has been created as the last step in preprocessing phase to prepare homophily network for clustering phase.For the demography information, the ages have been divided into 11 ranges (Table 1 and 2).

Community detection: City block distance has been used to find a distance matrix between two users (x and y) as follows:

$$\text{Distance} = \sum_{t=0}^{n} | xi - yi | \quad (2)$$

where X_i and Y_i represent the ratings that given for an item i by users x and y respectively.PAM algorithm has been applied on a distance metric to find communities of users who are similar in their taste of movies.As for the number of clusters,it is evaluated using (silhouettes plot).

The prediction model: The model has been designed depending on the theory of probability, The Naive Bayes model has been used as a prediction model shown in Eq. 3:

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+

Table 2: Compare our proposed with other methods (Cluster-based CF (Ju and Xu, 2013), A multi-level CF (Polatidis and Georgiadis, 2016) on the same first dataset 100K Movielens.

Length of Lists	The Method	Precision	Recall	F-measure
5	Proposed method	0.431	0.050	0.088
	A multi-level CF Polatidis and Georgiadis (2016)	0.050	0.060	0.055
10	Proposed method	0.384	0.112	0.147
	A multi-level CF Polatidis and Georgiadis (2016)	0.060	0.070	0.065
30	Proposed method	0.330	0.182	0.215
	Cluster-based CF Ju and Xu (2013)	0.098	0.393	0.157
40	Proposed method	0.295	0.234	0.229
	Cluster-based CF Ju and Xu (2013)	0.089	0.405	0.146
50	Proposed method	0.271	0.279	0.278
	Cluster-based CF Ju and Xu (2013)	0.078	0.417	0.131

$$P(CIL.A) = \frac{P(C),P(L.AIC}{P(A.C)} \quad (3)$$

In this phase, the model uses the age of user (A) and the last items that have been viewed by a user(L) as the prior knowledge to predict items (C).

RESULTS AND DISCUSSION

In this section, we present the results of the evaluation of the lists of recommendation and compare the result with the baselines and show the effect of using the modelling techniques (clustering methods) and demographic to enhance the prediction of recommendation system.The experimentsperformed on two movielens datasets with different sizes as shown in Fig. 3 and 4, respectively. Movielens dataset was gathered by the Grouplens research project at the university of Minnesota. The first data is 100K Movielens and the other data is 1M Movielens. The dataset is divided into training and test sets with percentage 70% and 30%, respectively.

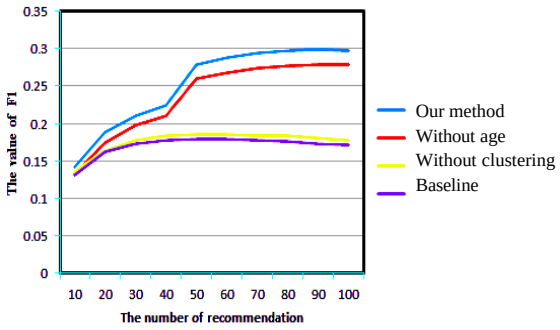

Fig. 3: The results of F-measure on the first Movielen dataset 100K

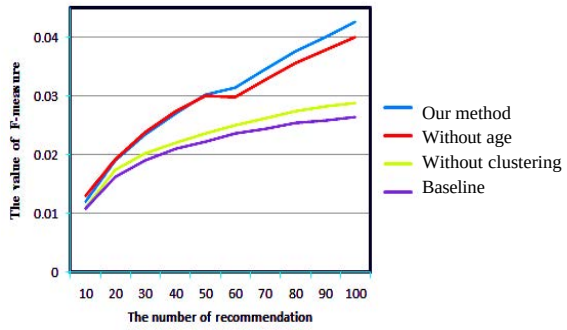

Fig. 4: The results of F-measure on the second Movielens dataset 1M

Three experiments have been performed to show the superiority of the proposed system over baseline as shown in the figures below.The lengths of the proposed recommended lists (top N) are from 10-100.

However, the proposed system has been evaluated using precision, recall to calculate the f-measure. Worth to mention, the values ofthese measures should not be understood as the absolute measure but these values use to compare with other algorithms with the same dataset (Cremonesi *et al*., 2008).

As clear, the proposed model has better performance over the baseline which represents one community without taking into our account the demographic information such as age in an account, Three experiments have been performed in terms f-measure. In the first experiment (green curve), the users have been considered as one community, i.e., the like-minded users in terms of preferences (homophily) is not taking into consideration. As for the demographic information such as age has been used as a factor prior consideration. The last experiment (blue curve) the communities which represent the like-minded users in terms of preferences and the age factor have been used.

In general, the performance of the proposed model (blue curve) is best when the length ofthe recommendation list increases for both datasets if compare it with baseline. Additionally, the performance of the proposed model is superior over the existing works. Important to say, the comparison with the other works has been constrained according to the published results in their researches.

CONCLUSION

In this study, we present two factors that can be affected on the prediction of recommendation system; The cluster model and demographic information. Clustering that makes the users in several communities instead of one community, where the people is different in terms of taste for movies which leads to forms communities in this regarding. Additionally, the age is considered as a good factor in determining the preferences. In concluding, It has been found that age factor in homophily communities improves the performance of the system.

REFERENCES

Ahn, H.J., 2008. A new similarity measure for collaborative filtering to alleviate the new user Cold-starting problem. Inform. Sci., 178: 37-51.

Beel, J., L.S. Nurnberger and M.A. Genzmehr, 2013. The Impact of Demographics (Age and Gender) and other User-Characteristics on Evaluating Recommender System. Springer, Berlin, Germany,.

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.

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.

Gogna, A. and A. Majumdar, 2015. A comprehensive recommender system model: Improving accuracy for both warm and cold start users. IEEE. Access, 3: 2803-2813.

Hofmann, T. and J. Puzicha, 1999. Latent class models for collaborative filtering. Proceedings of the 16th International Joint Conference in Artificial Intelligence, July 31-August 6, 1999, San Francisco, CA., USA., pp: 688-693.

Huang, X., Z. Qin and H.A. Chen, 2015. New user similarity measurement based on a local item space in collaborative filtering recommendation. J. Comput. Inf. Syst., 11: 3501-3508.

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.J. Rousseeuw, 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons, New York, ISBN: 0471878766, Pages: 342.

Miyahara, K. and M.J. Pazzani, 2000a. Collaborative Filtering with the simple Bayesian classifier. In: Pacific Rim International conference on artificial intelligence, Mizoguchi, R. and J. Slaney (Eds.). Springer, Berlin, Germany, ISBN: 978-3-540-44533-3, pp: 679-689.

Miyahara, K. and M.J. Pazzani, 2000b. Improvement of collaborative filtering with the simple Bayesian classifier. Inf. Technol. R. D. Center Mitsubishi Electr. Coporation, 2000: 679-689.

Murty, M.N. and V.S. Devi, 2011. Pattern Recognition: An Algorithmic Approach. Springer, Berlin, Germany, pp: 93-97.

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

Pillay, N., A.P. Engelbrecht, A. Abraham, M.C.D. Plessis and V. Snasel et al., 2015. Advances in nature and biologically inspired computing. Proc. World Congress Nature Biol. Inspired Comput., 2015: 39-41.

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 Systems Handbook. 2nd Edn., Springer, Berlin, Germany.

Shamri, M.Y.H.A. and N.H.A. Ashwal, 2014. Fuzzy-weighted similarity measures for memory-based collaborative recommender systems. J. Intell. Learn. Syst. Appl., 6: 1-10.

Sun, H.F., J.L. Chen, G. Yu, C.C. Liu and Y. Peng et al., 2012. JacUOD: A new similarity measurement for collaborative filtering. J. Comput. Sci. Technol., 27: 1252-1260.

Treerattanapitak, K. and C. Jaruskulchai, 2012. Exponential fuzzy C-means for collaborative filtering. J. Comput. Sci. Technol., 27: 567-576.

Wang, K. and Y. Tan, 2011. A New Collaborative Filtering Recommendation Approach Based on Naive Bayesian Method. In: International Conference in Swarm Intelligence, Ying T., S. Yuhui, Y. Chai and G. Wang (Eds.). Springer, Berlin, Germany, ISBN: 978-3-642-21524-7, pp: 218-227.