Handling User Cold Start Problem in Recommender Systems Using Fuzzy Clustering

Sugandha Gupta and Shivani Goel

Abstract Recommender engines have become extremely important in recent years because the count of people using Internet for diverse purposes is growing at an overwhelming speed. Different websites work on recommender systems using different techniques like content-based filtering, collaborative filtering, or hybrid filtering. Recommender engines face various challenges like scalability problem, cold start problem and sparsity issues. Cold start problem arises when there is no sufficient information for the user who has recently logon into the system and no proper recommendations can be made. This paper proposes a novel approach which applies fuzzy c-means clustering technique to address user cold start problem. Also, a comparison is made between fuzzy c-means clustering and the traditional k-means clustering method based on different set of users and thus it has been proved that the accuracy of fuzzy c-means approach is better than k-means for larger size of dataset.

Keywords Collaborative filtering · Cold start · Recommender system
Fuzzy clustering

1 Introduction

Recommender engines fall under the sub-category of information filtering systems that aim to forecast preferences or ratings given to the item by the user. With the developing technology, the influence of technology on everyone's life is increasing.

S. Gupta (✉) · S. Goel
Department of Computer Science and Engineering, Thapar University,
Patiala, India
e-mail: sugandhagupta_92@yahoo.in

S. Goel
e-mail: shivani@thapar.edu

D.K. Mishra et al. (eds.), *Information and Communication Technology for Sustainable Development*, Lecture Notes in Networks and Systems 10,
https://doi.org/10.1007/978-981-10-3920-1_15

Therefore, recommender engines are now-a-days an integral portion of E-commerce sites which helps in recommending items or products of interest to people all around the world. The major assistances of having a recommender systems are customer retention, information retrieval, personalization, and many more. Also recommender systems can be used on products such as music, books, restaurant, TV shows, and movies and presently are used in commercial websites successfully such as Movielens, Amazon, MovieFinder, ebay, Linkedln, Jinni, Facebook, and Myspace.

Basically, recommender systems compare the profile of a user to some basic characteristics and try to predict ratings given by a user to an item they had not yet well thought-out.

Recommender systems are categorized into the following three basic categories based on the way recommendations are generated:

- Collaborative Filtering Recommender Systems (also known as social filtering): The information is filtered by using the recommendations from different people. It works on the notion that people who agree with the evaluation of certain items or similar tastes or preferences in the past would agree in the future too (Linkedln[1]).
- Content-based Recommender Systems (also known to as cognitive filtering): It recommends items similar to those items which the user liked previously. Each item's content is symbolized by set of terms, usually the words that appear in a document. These terms represent the profiles of users, which are made after analyzing the contents of items seen by the user (Jinni[2]).
- Hybrid Recommender Systems: These use integration of two techniques, i.e., content-based filtering along with collaborative filtering which could be more effective in some cases.

Cold start problem, also known as new user problem or new item problem, is a special type of sparsity problem when a user or item has no ratings. Due to the lack of purchase history or rating information of a new user or item, it becomes a challenging task for recommender engines to generate appropriate recommendations for the new users and new items. Cold start problem is further divided into two categories

(1) New User cold start problem: It occurs when there are no ratings for the user who has arrived into the system.

[1] https://in.linkedin.com.

[2] http://www.jinni.com.

(2) New item cold start problem: It occurs when an item has just been added to the system and has not been rated as of yet.

The paper is divided into the following sections: Sect. 2 gives the details of previous work done on cold start problem. Section 3 gives the description of experiments applied to the user dataset and Sect. 4 gives the experimental results and their analysis. Section 5 concludes the paper.

2 Related Work

In the section described below, we review few prior studies linked to our suggested approach. Many researchers have done a plenty of work in the field of recommender engines. In this work, Adomavicius and Tuzhilin have given a view of collaborative, content, and hybrid recommender systems along with limitations of current recommendation techniques and their possible extensions [1]. The general concepts of collaborative filtering, their drawbacks and a clustering approach for huge datasets has been proposed by Sarwar et al. [2]. Sanchez et al. analyzed Pearson correlation metric and cosine metric together with the less common mean-squared difference in order to discover their advantages and disadvantages [3]. Also, Schein et al. have given a new evaluation metric known as the CROC curve and explained numerous components of testing strategies empirically in order to obtain better performance of recommender systems [4]. A comparative analysis of three different clustering methods, c-means clustering, k-means clustering, and SOM is proposed by Budayan et al. [5]. The work done by Ling Yanxiang et al. explains how to address cold start problem using character capture and clustering method [6]. Probabilistic neural network method to deal with cold start issues in the traditional collaborative filtering recommender engines is explained by Devi et al. [7]. Shaw et al. has given a technique which uses association rules to get rid of cold start problems in recommender systems [8]. Gupta and Patil proposed an efficient technique for recommender systems based on Chameleon Hierarchical clustering algorithm [9]. Ontology-based method to address cold start issue is proposed by Middleton et al. [10]. Son et al. proposed an application of fuzzy geographically clustering technique for handling the problem of cold start in recommender engines [11]. A novel hybrid approach to handle new item cold start issue in collaborative filtering uses both ratings and content information, as given by Sun et al. [12]. A new hybrid approach using ordered weighted averaging operator of exponential kind is used to get rid of cold start problem of recommender systems given by Basiri et al. [13]. A hybrid system for enhancing correlation using association rules and perceptron learning neural network to handle cold start issue in recommender systems was kept forward by Dang et al. [14].

3 Proposed Work

In this section, a sketch of our planned technique for addressing the user cold start problem is explained.

Our approach is to combine the two techniques in serial order one after the other. First we apply fuzzy clustering technique on the different attributes of user's demographic data which is entered by the user at the time of login into the system. Using the clustered data which we will get after applying fuzzy c-means clustering algorithm, we will apply Mysql using phpmyadmin for generating recommendations, aiming at new user recommendations. These clustering techniques help a new user to find a similar neighborhood. Thus recommendations generated are on the basis of the cluster the he is placed in. Top-N recommendations can be generated by querying in MySQL using aggregate analysis, i.e., by considering highest rating frequency or highest average rating for a movie. This would help to improve the quality of recommendations for a new user who enters the system.

The proposed system architecture is given in Fig. 1.

3.1 Clustering

Collaborative filtering recommender systems give best results when the user-item matrix is extensive and the dataset has high matching information according to the new user. The work done is based upon exploiting user's demographic data for finding similarity between the already existing user and the new user. Demographic information includes different user features like gender, occupation, religion, age, zip-code, race, locality, hobbies, marital-status, and many more. As compared to the existing k-means technique, fuzzy c-means clustering approach works by giving a membership value to every piece of data or data point equivalent to every cluster center based on the distance that lies between the cluster center and the data point. This technique allows one data point to be a part to two clusters or more. Therefore, if the data is nearer to cluster center, its association with specific cluster center is more.

As shown in the algorithm there are two important parameters a_{ij} and b_j. a_{ij} that represent association between ith data point and the jth cluster center and b_j represents the jth cluster center.

Algorithm:

Require: User demographic data D=$\{d_1, d_2, d_3, \ldots\ldots, d_n\}$, Set of cluster center C=$\{c_1, c_2, c_3, \ldots\ldots\ldots, c_m\}$. Assume fuzziness index ‘f’ $(1 \leq f \leq \infty)$ and Euclidean distance between the i^{th} data point and the j^{th} cluster center $\|d_i - b_j\|^2$

Ensure: New user groups: User Cluster

1. Arbitrarily select ‘m’ cluster centers.
2. Calculate fuzzy association between the data points and the cluster centers $'a'_{ij}$ using:

$$a_{ij} = 1 \Big/ \sum_{k=1}^{m} (e_{ij}/e_{ik})^{(2/f-1)}$$

3. Compute fuzzy centers of the clusters $'b'_j$ using:

$$b_j = \left(\sum_{i=1}^{n} (a_{ij})^f d_i\right) \Big/ \left(\sum_{i=1}^{n} (a_{ij})^f\right)$$

4. Repeat steps [2] and [3] till the objective function value O is minimized :

$$O(A, B) = \sum_{i=1}^{n} \sum_{j=1}^{m} (a_{ij})^f \|d_i - b_j\|^2$$

3.2 Recommendations

In Sect. 3.1, users are clustered into different groups or clusters by applying fuzzy c-means clustering approach based on their features (i.e., gender and age) which they will enter at the time of login into the system. Thus, we obtain the cluster’s recommendation to which the user belongs to by using the preprocessed data which will be stored in Mysql database for generating the recommendations. Therefore, the user gets the appropriate recommendations as we assume that the users with identical tastes shares same clusters and therefore tend to rate in the similar fashion.

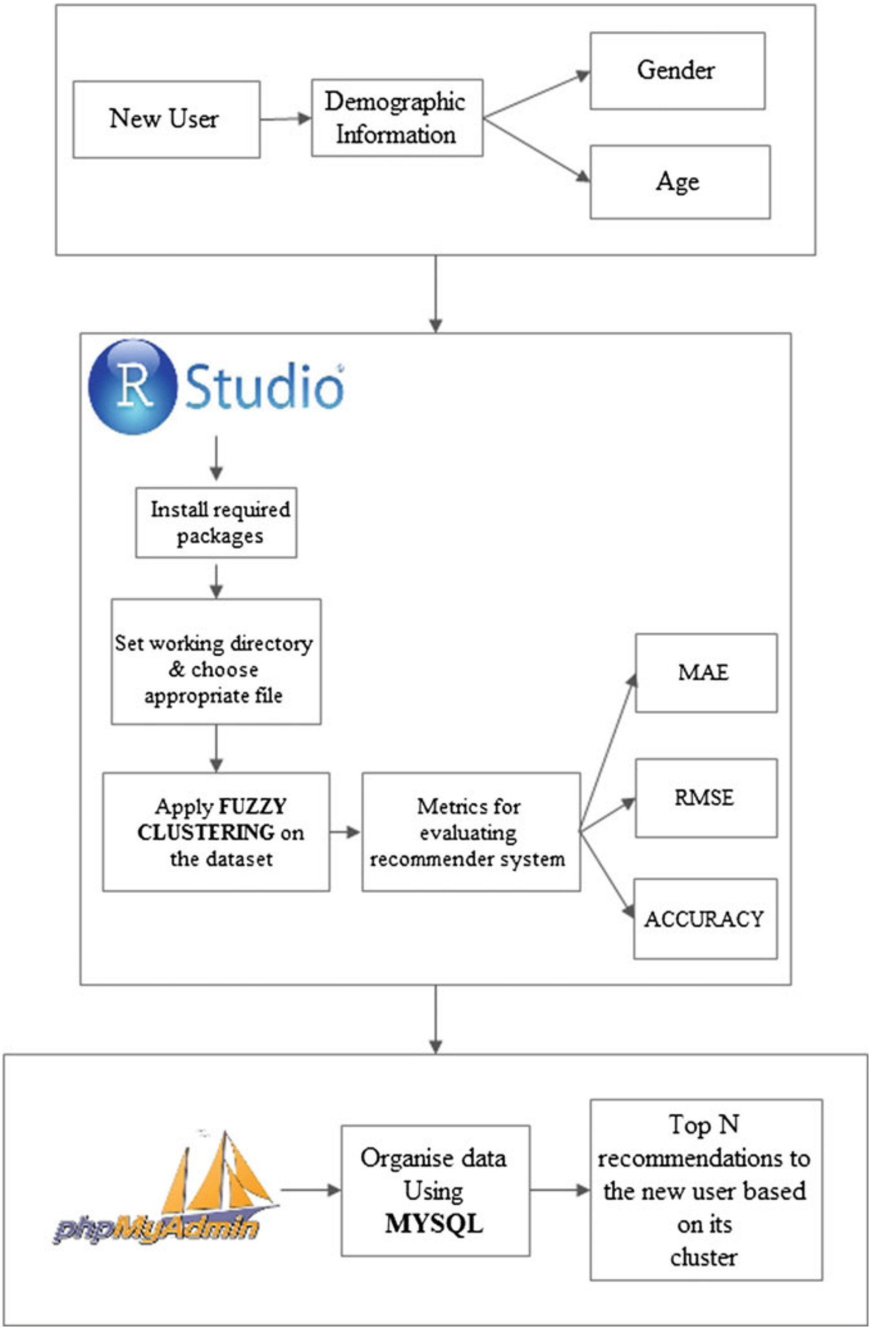
New User
Demographic Information
Gender
Age
R Studio
Install required packages
Set working directory & choose appropriate file
Apply FUZZY CLUSTERING on the dataset
Metrics for evaluating recommender system
MAE
RMSE
ACCURACY
phpMyAdmin
Organise data Using MYSQL
Top N recommendations to the new user based on its cluster

Fig. 1 Architecture of the collaborative filtering recommender system using fuzzy clustering technique

4 Experimental Results

4.1 Experimental Setup

To demonstrate the given approach, we apply it on MovieLens dataset, selecting 6040 users (where user features include userId, gender, occupation, and age) and their respective ratings. We preprocess the dataset using fuzzy c-means clustering technique which uses Euclidean distance to calculate similarity in Rstudio software and cluster the data using three clusters in order to evaluate our technique's performance on user cold start problem in collaborative filtering recommender systems. Therefore, a new user is recommended top-N recommendations according to his/her taste for which the preprocessed data is stored in Mysql database.

4.2 Evaluation Metrics

We have used three evaluation metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and accuracy for evaluating the predictions. Also, we compare the accuracy fuzzy c-means clustering approach with the traditional k-means clustering approach. The recommendation accuracy is better if we obtain lower values of MAE and RMSE.

$$MAE = mean(abs(x\$cluster - \bar{x})) \tag{1}$$

$$RMSE = sqrt(mean(x\$cluster - \bar{x}) \wedge 2) \tag{2}$$

$$Accuracy = mean(abs(x\$cluster - \bar{x}) \leq 1) \tag{3}$$

where $x\$cluster$ is predicted rating and $\bar{x}$ is value of actual rating.

4.3 Results

Since we have compared fuzzy c-means clustering approach with the traditional k-means clustering approach on the basis of evaluation metrics by varying the number of user present in the dataset, thus finding accuracy. The results demonstrate that as the number of users increase, fuzzy c-means technique comes out to be more accurate as compared to k-means approach. Therefore, it has been proved that fuzzy c-means approach is more accurate than k-means for large datasets as shown in Fig. 2.

The accuracy value for k-means clustering approach and fuzzy c-means clustering approach is shown is Table 1.

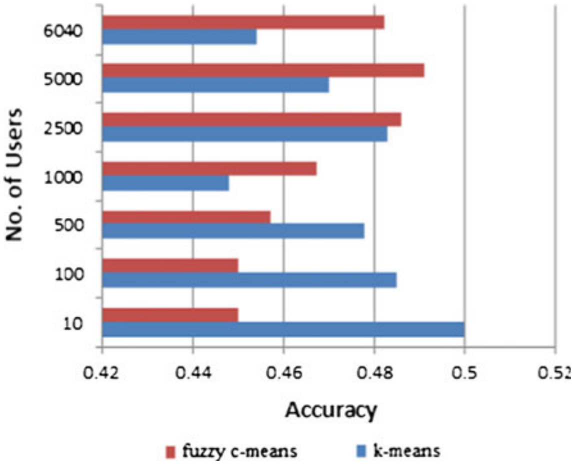

Fig. 2 Accuracy comparison of fuzzy c-means and k-means clustering approach

Table 1 Accuracy values for k-means and fuzzy c-means clustering for different number of users

Number of users	Accuracy	
	k-means	Fuzzy c-means
10	0.5	0.45
100	0.485	0.45
500	0.478	0.457
1000	0.448	0.467
2500	0.4828	0.4858
5000	0.4703	0.4912
6040	0.4542	0.4822

5 Conclusion and Future Scope

In this paper, a method has been proposed to address user cold start problem using fuzzy clustering technique. The basic idea behind our paper is to build clusters using fuzzy c-means clustering technique based on the attributes of the user. Hence, top-N recommendations for new users are successfully generated on the basis of the cluster he/she is placed into. The corresponding recommendation list is offered to the user (no matter old or new user) according to the cluster he/she is categorized into, based on the clustering algorithm used.

Also, a comparison is being made between fuzzy c-means clustering algorithm and the traditional k-means clustering algorithm on the basis of accuracy by varying number of users. It has been proved that as the number of users increase, fuzzy c-means clustering approach gives better accuracy as compared to the k-means clustering approach. The future scope will be to improve the accuracy of different clustering techniques by comparing it on different datasets.

References

1. Adomavicius G. & Tuzhilin A.: Toward The Next Generation of Recommender Systems: A Survey of the State-Of-The-Art and Possible Extensions: IEEE Transactions on Knowledge and Data Engineering, vol. 17, No. 6, pp. 734–749 (2005).
2. Sarwar B., Karypis J., Konstan J. & Riedl J.: Item Based Collaborative Filtering Recommendation Algorithms. In Proceedings of the 10^{th} International Conference on World Wide Web, ser WWW'01. pp. 285—295. ACM, USA (2001).
3. Sanchez J.L., Serradilla F., Martinez E. & Bobadilla J.: Choice of Metrics Used In Collaborative Filtering and Their Impact on Recommender Systems. In: Digital Ecosystems and Technologies: 2nd IEEE International Conference, pp. 432–436 (2008).
4. Schein A.I., Popescul A., Ungar L.H., & Pennock D.M.: Methods and Metrics for Cold-Start Recommendations. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '02, pp. 253–260. ACM, New York, USA (2002).
5. Budayan C., Dikmen I. & Birgonul M. T.: Comparing the Performance of Traditional Cluster Analysis, Self-organizing Maps and Fuzzy c-means Method for Strategic Grouping. In: Expert Systems with Applications, vol. 36, pp. 11772–11781 (2009).
6. Yanxiang L., Deke G., Fei C., & Honghui C.: User-based Clustering with Top-N Recommendation on Cold-Start Problem. In: Intelligent System Design and Engineering Applications (ISDEA), Third International Conference, pp. 1585–1589 (2013).
7. Devi M.K.K., Samy R.T., Kumar S.V. & Venkatesh P.: Probabilistic Neural Network Approach to Alleviate Sparsity and Cold Start Problems in Collaborative Recommender Systems. In: Computational Intelligence and Computing Research (ICCIC), IEEE International Conference, pp. 1–4 (2010).
8. Shaw G., Xu Y., & Geva S.: Using Association Rules to Solve the Cold-Start Problem in Recommender Systems. In: Lecture Notes in Computer Science, vol. 6118, pp. 340–347 (2010).
9. Gupta U. & Patil N.: Recommender System Based on Hierarchical Clustering Algorithm Chameleon. In: Advance Computing Conference (IACC), IEEE International, pp. 1006–1010 (2015).
10. Middleton S.E., Shadbolt N.R., & De Roure D.C.: Ontological User Profiling in Recommender Systems. In: ACM Trans. Inf. Syst., vol. 22, pp. 54–88 (2004).
11. Son L.H., Cuong K.M., Minh N.T.H., & Canh N.V.: An Application of Fuzzy Geographically Clustering for Solving the Cold-Start Problem in Recommender Systems. In: Soft Computing and Pattern Recognition (SoCPaR), International Conference pp. 44–49 (2013).
12. Sun D., Luo Z., & Zhang F.: A Novel Approach for Collaborative Filtering To Alleviate the New Item Cold-Start Problem. In: Communications and Information Technologies (ISCIT), 11th International Symposium, pp. 402–406 (2011).
13. Basiri J., Shakery A., Moshiri B., & Hayat M.Z.: Alleviating the Cold Start Problem of Recommender Systems Using a New Hybrid Approach. In: Telecommunications (IST), 5th International Symposium, pp. 962–967 (2010).
14. Dang T.T., Duong T.H., & Nguyen H.S.: A Hybrid Framework for Enhancing Correlation to Solve Cold-Start Problem in Recommender Systems. In: Computational Intelligence for Security and Defense Applications (CISDA), Seventh IEEE Symposium pp. 1–5 (2014).