A Hybrid Clustering Based Filtering Approach with Efficient Sequencing

Namita Mittal M.C. Govil Richi Nayak Rajesh Kumar Himanshu Gothwal Dwipayan Das

***Abstract*—A clustering technique is applied to integrate the contents of items into the item-based collaborative filtering framework in this paper. The information that is obtained from the clustering of content provides a way to introduce content information into collaborative recommendation and solves the cold start problem. The content-based filtering provides the content but may fail to understand what the user mean by that content while collaborative filtering may succeed in understanding the desire of user but may fail to bring the results for that request. However, we succeed to provide the relevant content to user by combining the content-based and collaborative filtering techniques and with the use of fuzzy C-means and the K-means clustering algorithm The results show that our approach contributes to the improvement of prediction quality of the item-based collaborative filtering, especially for the cold start problem.**

***Index Terms*—cluster, content based filtering, collaborative filtering, fuzzy C-means.**

I. INTRODUCTION

Whenever a user requests a search on a data set, it is very much desired that the system understands what user wants to retrieve. For this purpose the system has to use some filtering techniques for the request made. There are two widely used filtering techniques for this purpose, namely content-based filtering and collaborative filtering. Content-based filtering selects the right information for users by comparing representations of searching information to representations of contents of user profiles which express interests of users [3].

Manuscript received December 30, 2007.

Ms. Namita Mittal is working as Sr Lecturer, Department of Computer Engg, National Institute of Technology, Jaipur India. Phone: 91-141-2713324; fax: 91-141-2529028; e-mail: nmittal@ mnit.ac.in.

Dr M C Govil is with the Govt Mahila Engg College Ajmer, India as Principal. He is on leave from the National Institute of Technology Jaipur (e-mail: govilmc@yahoo.com).

Dr. Richi Nayak is with Queensland University of Technology, Australia. (email : r.nayak@qut.edu.au)

Dr. Rajesh Kumar is with National Institute of Technology, Jaipur working as Lecturer Electrical Engineering Department. (email:rkumar@gmail.com)

Himanshu Gothwal & Dwipayan Das are B.Tech (IT) students of National Institute of Technology Jaipur, India (email: hims333,dwipayan.das@gmail.com)

Content-based information filtering is generally used to search textual items relevant to a topic using techniques, such as Boolean queries, vector-space queries, probabilistic model, neural network and fuzzy set model [2]. However, content-based filtering has got two limitations. First it is hard for content-based filtering to provide unexpected or unforeseen recommendations, because all the information is selected and recommended is based on the content. Second it is difficult for new users to use content-based systems effectively because the user has not yet rated any movie and hence it's difficult for system to identify user's interest.

On the other hand collaborative filtering is the technique of using peer opinions to predict the interests of others. A target user is matched against the database to find group of users, who have historically had similar interests to target user [9]. Items that the users of group liked are then recommended to the target user [6]. Collaborative filtering has been applied in several systems like "The Tapestry text filtering system", developed by Nichols and others at the Xerox Palo Alto Research Center (PARC), The Group Lens project at the University of Minnesota etc. [14].

Collaborative filtering system overcomes limitations of content-based filtering but it has its own limitation. In traditional collaborative filtering approach, it is difficult for pure collaborative filtering to recommend a new item to user due to the absence of any user rating on the new item [5,7]. In this paper we propose to use the information from group cluster matrix to make predictions for a new item. Since the system suggests items to users based on the ratings of items, combined with contents of the items, the quality of recommendation is improved. In our experiment, it shows the results with proper relevance not only to the content but to the user interest also.

Section II gives work done in the field of collaborative filtering, Section III gives the overview of our proposed approach, Section IV explains the experiments done followed by the conclusion in Section V.

II. RELATED WORK

There are two main approaches for combining content-based and collaborative filters together. One approach is the linear combination of results of collaborative and content-based filtering, such as systems that are described by Claypool (1999) [4] and Wasfi (1999) [11]. ProfBuilder (Wasfi, 1999)[11] recommends web pages using both content-based and collaborative filters, and each creates a recommendation list without combining them. A hybrid approach is described by Claypool (1999) [4] for an online newspaper domain, combining the two predictions using an adaptive weighted average. The weight

of the collaborative component tends to increase as the number of users accessing an item increases. However, allocation of the weights of individual components of collaborative and content-based components is unclearly given by the [1]. The other approach is the sequential combination of content-based and collaborative filtering methods. In this system, firstly, content-based filtering algorithm is applied to find users, who share similar interests. Secondly, collaborative algorithm is applied to make predictions, such as RAAP [5] and Fab filtering systems[5].

RAAP is content-based collaborative information filtering for helping the user to classify domain specific information found in the Web, and also recommends these URLs to other users with similar interests. Similar interests of user are determined using the ratings of content. Fab System uses content based techniques instead of user ratings to create profiles of users. So the quality of predictions is fully dependent on the content-based techniques. Sometimes it results into poor predictions as the inaccurate profiles result in inaccurate correlations with other users. Similarity in a collaborative filtering is done by two main techniques which are item-based and user-based. Sarwar [11] has proved that item-based collaborative filtering is better than user-based collaborative filtering at precision and computation complexity.

III. OVERVIEW OF THE APPROACH

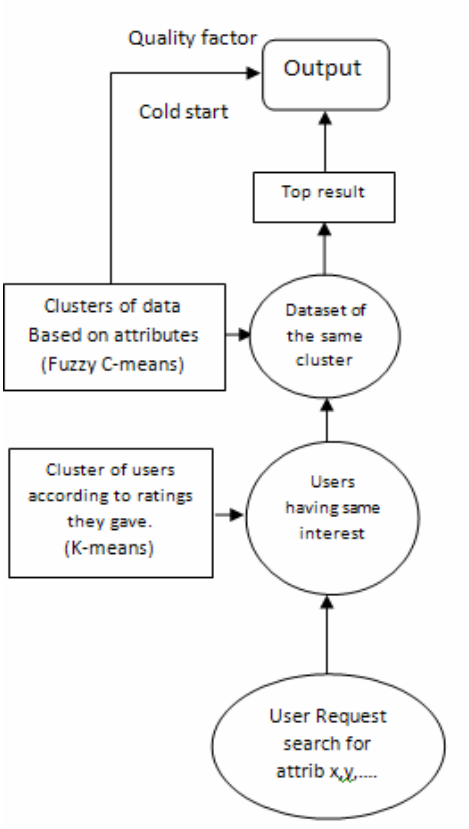

Figure 1: Overview of the approach

In this paper, we suggest a technique that combines user interest with the content of the data to improve its prediction quality and solve the cold start problem. Hence we use two clustering based models *ICHM* and *UCHM* [1]. Fuzzy C-means clustering is applied for *ICHM* and K-means clustering is used for *UCHM*. The detailed procedure of our approach, in Fig.1, is described as follows:

When a user request for a search of specific content, the following operations are performed:

1) Find the users of similar interest by k-means clustering algorithm [6].
2) Find the content (products/items) that the users of the same cluster have rated high on the scale.
3) Add the content requested in the list of contents and then apply fuzzy c-means algorithm [8] to find contents in the same cluster as that of the content requested.
4) Find the common set between the two sets collected during steps (2) and (3). This set represents the best results for the search.
5) To resolve the cold start problem set a threshold percentage(quality factor) for the content-based clustering, find the content which has percentage membership greater than the quality percentage in the set collected during step (3) and include the result in the search result
6) Remove all the redundancies.

A. Fuzzy C-means Clustering

Given a finite set of data *X*, the problem of clustering in *X* is to find several cluster centers that can properly characterize relevant classes of *X*. In classical cluster analysis, these classes are required to form a partition of *X*, such that the degree of association is very strong for data within blocks of partition and very weak for data in different blocks. However this requirement is too strong in many practical applications and it is thus desirable to replace it with a weaker requirement. When the requirement of a crisp partition is replaced with a weaker requirement of a fuzzy partition on *X*, we call the emerging grouping method as fuzzy clustering [13].

Fuzzy C-Means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters[15]. C represents the number of clusters. The FCM employs fuzzy partitioning such that a data point can belong to all groups with different membership grades between 0 and 1.

FCM is an iterative algorithm. The aim of FCM is to find cluster centers (centroids) that minimize a dissimilarity function. The algorithm is based on the assumption that the desired number of clusters c is given and, in addition, a particular distance, a real number m Є (1, ∞), and a small positive number ε , serving as a stopping criterion, are chosen[13].

IV. EXPERIMENT

Considering the contents as shown in Table 1 where C1 and C2 are attributes of the data set of movies M for example the year and the Genre of movie where numeric values are code for keywords as shown in Table 1.

TABLE1: Data and content of data

	M 1	M 2	M 3	M 4	M 5	M 6	M 7	M 8	M 9	M 10	M 11	M 12	M 13	M 14	M 15
C 1	1	3	1	2	2	3	1	2	1	3	1	1	2	3	1
C 2	1	3	3	2	3	2	3	1	3	1	1	3	3	3	3

TABLE 1a: Meaning of codeword

	1	2	3
C1	70s	80s	90s
C2	Action	Drama	Comedy

And a sample of user ratings are as shown in Table 2 where ten users (U1-U10) have rated 12 movies (M1-M12) out of the above 15 on the scale of 5.

TABLE 2: user ratings on movies

	M 1	M 2	M 3	M 4	M 5	M 6	M 7	M 8	M 9	M 10	M 11	M 12
U1	5	4	1	1	5	1	3	4	4	1	3	2
U2	2	4	2	4	1	4	5	2	5	1	3	2
U3	4	5	5	3	4	2	5	5	3	5	3	4
U4	3	4	1	5	2	3	4	3	5	1	2	4
U5	5	1	1	3	5	1	5	2	1	2	3	3
U6	4	3	2	3	3	4	4	4	5	4	2	3
U7	3	5	1	5	4	2	2	3	2	2	3	4
U8	1	5	4	3	3	5	2	3	2	3	4	1
U9	5	3	2	2	2	5	2	4	5	1	3	4
U 10	3	5	1	4	1	3	3	4	4	5	4	1

Taking the case that the user U3 requested for a content (1,3) i.e comedy movie of 70s. , the results of collaborative filter, content based filter and the Hybrid filter are shown in the next section.

B. Collaborative filtering

When we use collaborative filter only then the clusters of users obtained as outcome are shown in Table 3 where the Ten users (U1-U10) those rated the movies above, have been divided into five clusters using K-means clustering algorithm.

TABLE3: User and cluster to which they belong

User ->	1	2	3	4	5	6	7	8	9	10
Cluster No.	5	4	1	3	1	1	3	5	3	2

Users in same cluster are U3, U5 and U6 since none of them have rated the 15th data D15 i.e (1,3) of Table 1, so it will not be included in the results although it matches with the content. That is a cold start problem, faced by the collaborative filter.

C. Content search based filtering

When we use content search based filtering only, then appending the table of contents with (1,3) i.e a new comedy movie of 70s and afterwards applying the Fuzzy C-Means clustering algorithm, the following percentage is obtained which shows belongingness of data set in two clusters CL1 and CL2:

TABLE 4: Belongingness of data to cluster CL1 and CL2

	M1	M2	M3	M4	M5	M6	M7	M8
CL1	34.1	45.7	97.6	13.8	79.5	15.5	97.6	9.5
CL2	65.9	54.3	2.4	86.2	20.5	84.5	2.4	90.5
	M9	M10	M11	M12	M13	M14	M15	M16
CL1	97.6	12.1	34.1	97.6	79.5	45.7	97.6	97.6
CL2	2.4	87.9	65.9	2.4	20.5	54.3	2.4	2.4

The filter will show results in terms of literal matching. Result will be M3, M5, M7, M9, M12, M13, and M15. Note that M2 match to the user interest according to TABLE 2 but is not included in the results. However, it is not confirmed that it matches with the interest of the user or not. Hence the result of content based filter may not be in the sequence desired by the user.

D. Hybrid filtering

Firstly using the collaborative filtering, users in same cluster are U3, U5, and U6 as shown in Table 3. Now applying the content based filtering on the movies that user U3, U5 and U6 rated high contains movies M3, M9, M7, M12 and M2. But other movies that have a high percentage (Table 4) but not rated should also be included which are M13 and M15 using a quality factor of 75% membership. The result set of our experiment obtained is M3, M9, M7, M12, M2, M13, and M15 which is organized according to order user required to have.

V. CONCLUSION

The use of K-means for collaborative filtering followed by Fuzzy C-means clustering for content filtering gives a deep filtering to what the user exactly needs. Double filtering of this hybrid model increases the probability of getting the results desired by user. The proposed algorithm performs better as the cold start problem is solved by the recent request /implicit interest of user but not by the interest from user profile/explicit interest which was created at the time of registering the user.

REFERENCES

[1] Qing Li, Byeong Man Kim, 2003. *An approach for combining content based and collaborative filtering*, In proc. of the sixth international workshop on information retrieval with Asian languages, page 17-24

[2] Anick, P. G., Brennan, J. D., Flynn, R. A., Hanssen, D. R., Alvey, B. and Robbins, J.M.. 1990. *A Direct Manipulation Interface for Boolean Information Retrieval via Natural Language Query*, In Proc. ACMSIGIR Conf., pp.135-150.

[3] Balabanovic, M. and Shoham, Y. 1997. *Fab: Content-Based, Collaborative Recommendation*, Communications of the ACM, 40(3), pp.66-72.

[4] Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D. and Sartin, M. 1999. *Combining content based and collaborative filters in an online newspaper* In Proc. ACM-SIGIR Workshop on Recommender Systems: Algorithms and Evaluation.

[5] Delgado, J., Ishii, N. and Ura, T. 1998. *Content-based Collaborative Information Filtering: Actively Learning to Classify and Recommend Documents*, In Proc.Second Int. Workshop, CIA'98, pp.206-215.

[6] Han, J., and Kamber, M. 2000. *Data mining: Concepts and Techniques*. New York: Morgan-Kaufman.

[7] Herlocker, J., Konstan, J., Borchers A., and Riedl, J.. 1999. *An algorithmic framework for performing collaborative Filtering*, In Proc. ACM-SIGIR Conf., 1999, pp. 230-237.

[8] Oard, D.W. and Marchionini, G. 1996. *A conceptual framework for text filtering*, Technical Report EETR- 96-25, CAR-TR-830, CS-TR3643. Ogawa, Y., Morita, T. and Kobayashi, K. 1991. *A fuzzy document retrieval system using the keyword connection matrix and a learning method*, Fuzzy sets and Systems, 1991, pp.39, pp.163-179.

[9] O'Conner, M. and Herlocker, J. 1999. *Clustering items for collaborative filtering*, In Proc. ACM-SIGIR Workshop on Recommender Systems.

[10] Sarwar, B. M., Karypis, G., Konstan, J. A. and Riedl, J..2001. *Item-based Collaborative Filtering Recommendation Algorithms*, In Proc. Tenth Int. WWW Conf. 2001, pp. 285-295.

[11] Wasfi, A. M. A.. 1999. *Collecting User Access Patterns for Building user Profiles and Collaborative Filtering*, In Int. Conf. on Intelligent User Interfaces. pp.57- 64

[12] Hauver, D. B.. 2001. *Flycasting: Using Collaborative Filtering to Generate a Play list for Online Radio*, In Int. Conf. on Web Delivery of Music.

[13] J.C.Bezdec, 1981, *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum Press, New York.

[14] Resnick, P., Iacovou, N., Suchak, M., Bergstorm, P. and Riedl, J.. 1994. *GroupLens: An open architecture for collaborative filtering of Netnews*, In Proc. ACM. Conf. on Computer-Supported Cooperative Work pp.175-186..

[15] J. C. Dunn (1973): "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", *Journal of Cybernetics* 3: 32-57