A Hybrid Recommender System for Dynamic Web Users

Shiva Nadi
Department of Computer Engineering, Islamic Azad University of Najafabad Isfahan, Iran

Mohammad Hossein Saraee
Department of Electrical and Computer Engineering, Isfahan University of Technology Isfahan, Iran

Ayoub Bagheri
Department of Electrical and Computer Engineering, Isfahan University of Technology Isfahan, Iran

Abstract

Nowadays, providing tools that eases the interaction of users with websites is a big challenge in e-commerce. Recommender systems are useful tools which adapts the environment of websites compatible with users needs. In this paper, applying a hybrid collaboration and content based technique a model for recommendation system is proposed. Presented model works in two offline and online phases. In offline step the behavior of users' models with a combined FCM and ant based clustering algorithm and in online step suitable recommendations extracts for presenting to active user. The model is implemented and tested as a recommender system for personalizing website of "Information and Communication Technology Center" of Isfahan municipality in Iran. The results shown are promising and proved that applying more efficient clustering technique for modeling users behavior provide us with more interesting and useful patterns which consequently making the recommender system more functional and robust.

1. Introduction

With the rapid growth of information available on the Web and increasing needs for easy use of web contents, using websites that are compatible with user's preferences is much raised. Web personalization is a process in which Web information space adapts with user's interests [8]. Web usage mining techniques are widely employed for extracting knowledge about user interests [6]. However, as the vagueness and imprecision in user interests are key features of web personalization systems (WPS), traditional models which use hard computing techniques (0 and 1) are inadequate. Since, the navigational behavior of users on the web is uncertain, using fuzzy clustering algorithms are more suitable for usage mining.

In the recent years many researches tries to personalize websites. Castellano et al. have presented an approach for expressing similarity between Web users according to their access patterns. They considered the access time to a web page as a parameter in the model [4]. Castellano et al. have also used a neuro-fuzzy model to develop a Website personalization framework for dynamic suggestion of URLs to users [5]. A different approach is purposed in the work of Kraft et al. in which they extracted fuzzy rules from user profiles and used them for information retrieval. Web user profiles usually reflect only access behavior of users and not keywords on a website. In this work, they rated web pages for building user interest profiles [3]. A bi clustering approach to correlate web users and pages are purpose in the work of [12], their purposed approach is a three step process which concentrates on the principles of spectral clustering analysis and provides a fuzzy relational scheme for both user and page clusters but they didn't consider to active users for providing dynamic recommendation. Another approach for combination of content mining and web usage mining in order to predicting user navigational behavior is represented by [11]. The frequent word sequences are used for capturing textual content of web pages. Log files data are combined with content clusters for extracting rules about user's behavior. Their work is improved in this work by employing fuzzy techniques for user clustering and predicting rated recommendations for active user.

In this paper we propose a model applying both user and URL clustering using fuzzy techniques for better dynamic recommendation process. The model is implemented and tested for the website of "Information and Communication Technology Center" of Isfahan municipality in Iran which verifies that fuzzy clustering approach in our model can lead to a better web personalization and recommendation.

The rest of the paper is organized as following. In section 2 the architecture of the model is described and in subsections the model's strategy for

knowledge discovery and dynamic recommendation process is represented. Experimental results about the implemented system are illustrated in section 3. A brief discussion and evaluation about model is represented in section 4 and finally in section 5 the conclusion and future research directions are presented.

2. Architecture

The architecture of the web personalization model is represented in Figure1. Web server transparently provides user with the personalized environment. When user requests an URL, the server retrieves and returns the requested URL followed by a list of URLs which may be interested to the user.

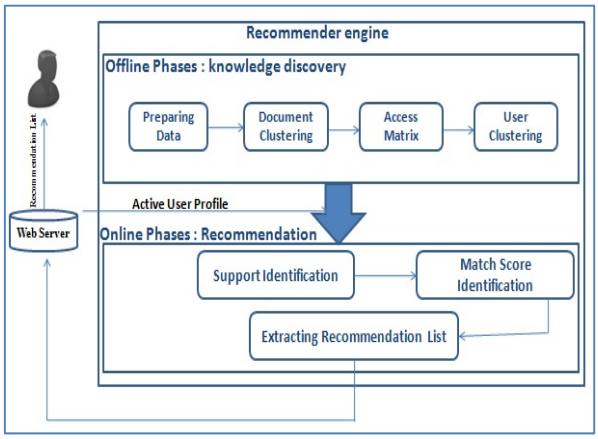

Figure 1. Architecture of the Proposed Model

As illustrated in Figure 1, two major steps of a web personalization system are knowledge discovery and recommendation. In first step, user preferences are identified using web access log data called web usage data (WUD). In the next step, the achieved knowledge is used to identify the possibly interested URLs to the users. This recommendation can be done in different ways such as adding related hyperlinks to the last web page requested by the user [2]. In this paper, Represented method dynamically recommends the highest match score URLs to active user. Represented method steps are described in the following sections:

2.1 Knowledge discovery Process

In this step, using the previous behavior of each user, knowledge that represents user's preferences is extracted. For extracting knowledge from web usage data, following steps will be performed:

2.1.1 Preprocessing

Log files have useful information about access of all users to a specific website. Extracting these information, reformatted log file which contains useful information such as "time, date, accessed URL and IP address" is formed and useless requests such as accesses to images are removed from log file in data cleaning process.

2.1.2 Document Clustering

In this work, we used content mining approach for document clustering. Assume $P = \{p_1, p_2, \dots, p_k\}$ is the set of k website's pages. Applying the clustering algorithm shown in Figure 2, pages were grouped in content based clusters.

1. *Clear each document from stop words such as: about, all, am, almost, as, be, by, but, do and any other word which haven't any key role in determining the content of document.*
2. *Identify document keywords by TF-IDF technique.*
3. *Assign each document keyword list as a document to a single cluster.*
4. *Merge primary clusters based on the Jaccard coefficient similarity measure. Defined as:*

$$sim(p_x, p_y) = \frac{|p_x \cap p_y|}{|p_x \cup p_y|} \qquad (1)$$

$|p_x \cap p_y|$ *Represents the number of common words and* $|p_x \cup p_y|$ *represents total number of words between two basic clusters.*

5. *The second step repeated until all documents being clustered into a pre defined number of clusters.*
   $DC = \{DC1, DC2, \dots, DC_n\}$ *is the result set. Each* DC_i *represents a set of URLs with similar content.*

Figure 2. Document clustering Steps

2.1.3 Integrating access matrix

In this step, the previously obtained document clusters are merged with reformatted log file and according to the result, access table can be produced. Then, users are clustered based on their behavior in access to document clusters. Assume U= {U1, U2, …, Um} as a set of users where each user is identified by distinct IP address and A= { A1,A2 , … ,Am } is a set of user's accesses to document clusters. For example A_k indicates the access list of user, *k*, to a subset of document clusters ($A_k \in$ DC, for K=1 to n). Access table is represented as a matrix with *m* rows and *n* columns where each entry represents the interest degree of *i-th* user to *j-th* document cluster. We use the interest degree which proposed by Castellano et al., They defined it as the ratio of the number of accesses to each document cluster to the total number of accesses to all document clusters for each user [6].

$$\boldsymbol{Val}(\boldsymbol{DC_i}, \boldsymbol{U_i}) = \frac{|\{\boldsymbol{DC_j}|\boldsymbol{DC_j} \in \boldsymbol{A_i}\}|}{|\boldsymbol{A_i}|} \qquad (2)$$

Calculated access matrix normalizes in the range of 0 to 1 and is the input data for clustering algorithm.

2.1.4 Fuzzy Ant based User Clustering

According to access matrix, users with similar interests can be clustered together. In this work, we used a two phase algorithm. This compound algorithm results in UC= $\{UC_1, UC_2, \ldots, UC_n\}$ where each UC_i represents a set of users with similar interesting patterns.

The ant based algorithm provides acceptable clusters of data without any knowledge of the initial clusters. In the ant based algorithm if an object is assigned to an inappropriate heap then it takes long time to be transported to a better heap. Overcoming this problem, a synthetic algorithm proposed by Kanade and Hall (2003) that uses the integrated FCM and ant based clustering algorithms [10]. In the web recommendation field we supposed that objects are users associated with a vector of fuzzy numbers which indicates their interests to document clusters.

The fuzzy C-means algorithm requires good initialization. These initial values are provided by the ant based algorithm. The result will be small homogenous heaps that will be merged by repeating the steps. By increasing the number of iterations the number of heaps decreases. Figure 3 shows the hybrid algorithm used in this study to cluster users in appropriate groups [10]:

6. *Scatter the users randomly on the board*
7. *Initialize the ants with random position, and random direction*
8. *For N iterations do*
    a. *For each ant do*
        i. *Move the ant*
        ii. *If the ant is carrying an user U then possibly drop the user U else*
        iii. *Possibly pick up a user*
9. *Use the cluster centers obtained in step 3 to initialize cluster centers for the fuzzy C-means algorithm*
10. *Cluster the data using the fuzzy C-means algorithm*
11. *Harden the data obtained from the Fuzzy C-means algorithm, using the maximum membership criterion, to form new heaps*
12. *Repeat step 1-6 by considering each heap as a single object*

Figure3: User clustering algorithm

The advantage of this algorithm lies within the calculation of initial clusters' centers which is based on Ant colony algorithm.

2.2. Recommendation Process

When a new user starts a transaction, our model matches the new user with the most similar user clusters and provides suitable recommendations for him/her.

2.2.1 Support Identification

When a new user starts a transaction, our model matches the new user with the most similar user clusters and provides suitable recommendations to him. Thakur et al. represented a criterion called support value which was calculated through following steps [2]:

Step1. Assign active user to a new cluster.

UC_{new}= { U_{new} }.

Step2. Complete access matrix for active user according to the equation 2.

Step3. Calculate support value of UC_{new} to existing user clusters UC_i using Equation 3.

$$support(UC_{new}, UC_i) = \frac{\sum |user(val(DC, UC_{new}) - val(DC, UC_i))|}{UC_i \cup UC_{new}} \quad (3)$$

$val(DC, UC_i)$ shows the interesting value of users in UC_i to the documents in selected DC.

2.2.2 Match Score Identification

The match score calculation defines highest match user cluster for active user. Match score criterion is represented by Thakur et al. (2009). This parameter between UC_{new} and UC_i is defined as following [2].

$$match(UC_{new}, UC_i) = 1 - support\ (UC_{new}, UC_i) \quad (4)$$

This give us a list of corresponding user clusters from the highest match score down to lowest match score.

3. Experimental results

In this work, the log file is given from "Information and Communication Technology Center" of Isfahan municipality in Iran (FAVA) for IP address 80.191.136.6. We collected log file during a period of one week. After data cleaning in preprocessing step, the number of requests was 52322 which were structured to 12332 sessions. The number of accessed URLs in this website was 200 pages. Employing content based document clustering algorithm we grouped the URLs to 5 clusters. Then user's behavior modeled as access matrix. Using equation 2, interest degrees are calculated as a 12332×5 matrix which is shown in Table 1, this is the input of user clustering algorithm.

Table 1. Access Table

IP	DC1	DC2	DC3	DC4	DC5
66.55	0.6	0	0.4	0	0
78.39	0.3	0	0.3	0	0.3
66.249	0.2	0.2	0	0.5	0.1

In next step, applying FCM and compound ant based clustering algorithm on access table, 10 user clusters was gained.

The center of the cluster is a vector, which is computed as the mean of user preferences in the user cluster. Table 2 indicates the center of clusters 1.

Table 2. Center of cluster

UC1	DC1	DC2	DC3	DC4	DC5
U112	0.7	0	0.3	0	0
U21	0	0.2	0	0.3	0.6
U39	0.5	0.1	0.1	0	0.2
U14	0.2	0.5	0.3	0	0
Center	0.37	0.20	0.17	0.07	0.20

Next, we considered an entry of log file as a new user who is recently connected to the website; access table is calculated again for measuring the interest degree of new user to document clusters. Table 3 shows parts of this matrix.

Table 3. Reformatted access table for new user

Center	DC1	DC2	DC3	DC4	DC5
UC1	0.37	0.20	0.17	0.07	0.20
UC2	0.0	0.25	0.0	0.51	0.24
UC3	0.43	0.0	0.10	0.30	0.0
UC4	0.0	0.50	0.30	0.0	0.20
UCnew	0.03	0.0	0.0	0.53	0.44

According to equation 3, support value of active user for each user cluster is calculated and shown in Table 4 and using equation 4, match score tables are calculated as showed in Table 5.

Table 4. Support and Match score calculation for UCnew

Uclusters	Support	Match	Rate
UC1	0.7	0.3	L
UC2	0.26	0.74	VH
UC3	0.63	0.37	M
UC4	0.56	0.44	H

Here, the new user has highest match score to the user cluster number 2. According to table 3 users in this user cluster have shown most interest to documents in DC4, this document cluster is chosen for making recommendations for new user.

4. Discussion

Our recently work has been improved in this paper by employing a hybrid clustering algorithm for user clustering and supporting active user by a range of rated relevant recommendations. The results of

this content based recommender system has been successfully modeled in this paper. Suggested recommendation model combines item and user clustering employing benefits of both collaborative and content based techniques. Applying model on the access log file of FAVA website, a set of rated items was derived from user preferences for dynamic recommendation.

The advantages of the proposed model are summarized as following. Firstly, this model acts effectively in identifying user preferences.

One common limitation in clustering algorithms is defining a suitable number of clusters which is solved with using a hybrid FCM and ant based clustering algorithm. Otherwise, uncertainty among user interests is an important issue which is clearly considered in user clustering. Additionally, this approach provides dynamic user clustering for active users and supports them with personalized environment. Moreover, any changes in website documents can be simply informed to all group users which are highly interested on that page. While the advantages of the model are clear, there are some limitations in this research area that we consider them as future research directions. Document clustering and user clustering in this approach are considered in separate processes which is time consuming. In order to solve this problem a Two-way clustering method may be used. This strategy in the same time not only clusters the objects but also the features of the objects will be clustered. However, the proposed architecture provides us with reasonable results. Evaluating the effectiveness of the purposed model, the top n recommendations are evaluated using precision, recall and F1 measures. Precision measure shows the accuracy of presented recommendations and recall is a measure of completeness and F1 measure combines Precision and Recall and is the harmonic mean of precision and recall. In Figure 4, 5 and 6 these metrics are evaluated for FARS recommender system which uses hybrid clustering method with systems which uses ant based clustering [9] and FCM [7] .

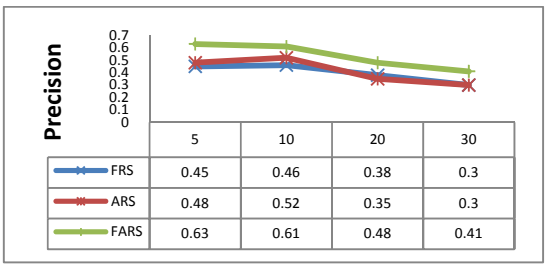

Figure 4. Precision measures for N top recommendations

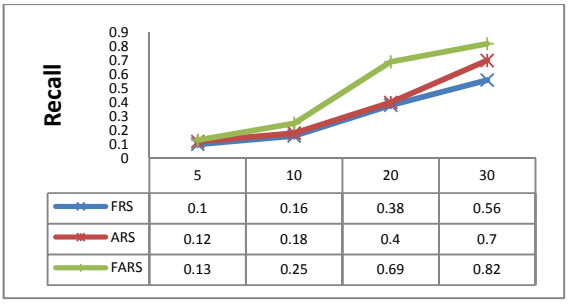

Figure 5. Recall measures for N top recommendations

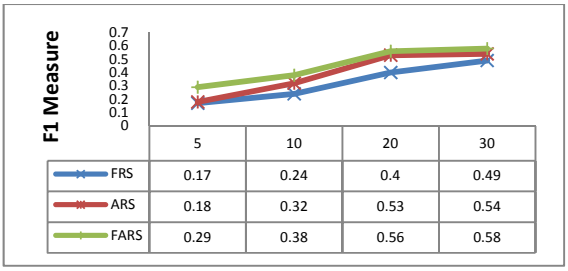

Figure 6. F1 measures for N top recommendations

The Figures indicates the proposed approach improves the performance of the models which uses FCM and ant based clustering methods. Therefore our approach shows more qualified recommendations.

5. Conclusions

In this paper, a model for dynamic recommendation based on a hybrid clustering algorithm is proposed. This model analysis the users behaviors and depend on the interests of similar patterns provides appropriate recommendations for active user. The model uses the benefits of both content based and collaborative based recommender systems. The results of evaluations shows that using more efficient algorithms for finding similar users lead to recommender system that provides more interesting recommendations for website users. Our work can be extended by considering the effect of users' feedback for increasing the quality of recommendation.

6. References

[1] B. Goethals, "Survey on Frequent Pattern Mining", P.O. box 26, FIN-00014 Helsinki Finland, 2003.

[2] B. K Thakur, S.Q Abbas, A.K Trivedi, "A Recommender System to Personalize the Environment of Web User", IEEE International Advance Computing Conference Patiala, India, 2009.

[3] D. H.Kraft, J. Chen, M. Bautista, M.J., and M.A. Vila, "Textual Information Retrieval with User Profiles Using Fuzzy Clustering and," Intelligent Exploration of the Web, Heidelberg, Germany: Physica-Verlag, 2002.

[4] G. Castellano, A. M. Fanelli, C. Mencar and M. Alessandra Torsello, "Similarity-based Fuzzy clustering for user profiling", IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology – Workshops, 2007.

[5] G. Castellano, A.M. Fanelli, P. Plantamura, M.A. Torsello, A Neuro-Fuzzy Strategy for Web Personalization, Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, 2008.

[6] M. P. Singh, "Web Usage Mining and Personalization", a chapter in Practical Handbook of Internet Computing, Munindar P. Singh (ed.), CRC Press, 2005.

[7] Sh. Nadi, M.H. Saraee, M. Davarpanah "A Fuzzy Recommender System for Dynamic Prediction of User's Behavior", 5th International Conference for Internet Technology and Secured Transactions, 2010.

[8] O. Nasraoui, "World Wide Web Personalization", in J. Wang (ed), Encyclopedia of Data Mining and Data Warehousing, Idea Group, 2005.

[9] P. Bedi., Sharma R. and Kaur H., Recommender System based on collaborative behavior of ants, 2009, Journal of artificial intelligence,ISSN 1994-5450, PP. 40-55

[10] Kanade P. M. and L. O. Hall, "Fuzzy ants as a clustering concept". North American Fuzzy Information Processing Society, NAFIPS 2003,22nd International Conference of the, pp. 227–232, 2003.

[11] S. Taherizadeh and N. Moghadam, 2009. Integrating web content mining into web usage mining for finding patterns and predicting users behaviors, *International Journal of Information Science and Management*, Vol, 7, No.1, pp.51-66

[12] V.A. Koutsonikola and A.I. Akali, 2009. A fuzzy bi clustering approach to correlate web users nad web pages, Int.J.Knowledge and Web Intelligence, Vol.1, Nos.1/2, pp. 3-23

