

A fuzzy co-clustering approach for hybrid recommender systems

Rana Forsati^{a,*}, Hanieh Mohammadi Doustdar^b, Mehrnosh Shamsfard^a, Andisheh Keikha^a and Mohammad Reza Meybodici^c

^a*NLP Research Lab, Faculty of Electrical and Computer Engineering, Shahid Beheshti University, Tehran, Iran*

^b*Department of Computer Engineering, Islamic Azad University, Qazvin Branch, Qazvin, Iran*

^c*Department of Computer Engineering, Amirkabir University of Technology, Tehran, Iran*

Abstract. Many efforts have been done to tackle the problem of information abundance in the World Wide Web. Growth in the number of web users and the necessity of making the information available on the web make web recommender systems very critical and popular. Recommender systems use the knowledge obtained through the analysis of users' navigational behavior to customize a web site to the needs of each particular user or set of users. Most of the existing recommender systems use either content-based or collaborative filtering approach. It is difficult to decide which one of these approaches is the most effective one to be used, as each of them has both strengths and weaknesses. Therefore, a combination of different methods as a hybrid system can overcome these limitations and increase the effectiveness of the system. This paper introduces a new hybrid recommender system by exploiting a combination of collaborative filtering and content-based approaches in a way that resolves the drawbacks of each approach and makes a great improvement in the variety of recommendations in comparison to each individual approach. We introduce a new fuzzy clustering approach based on genetic algorithms and create a two-layer graph. After applying this clustering algorithm to both layers of the graph, we compute the similarity between web pages and users, and propose recommendations using the content-based, collaborative and hybrid approaches. A detailed comparison on all the mentioned approaches shows that the hybrid approach recommends the web pages which haven't been yet viewed by any user, more accurately and precisely than other approaches. Therefore, the evaluation of the results reveals that the novel proposed combination approach achieves more accurate predictions and more appropriate recommendations than each individual approach.

Keywords: Content-based approach, collaborative filtering approach, hybrid approach, fuzzy clustering, two-layer graph

1. Introduction

The World Wide Web has become one of the most important communication tool and information retrieval source. Due to massive influx of information on the Web, it is difficult to find the useful information among distributed information sources. Admittedly, it is essential to predict the users' needs in order to improve the usability and user retention of a web site. Recommender systems are proposed to fulfill this aim

in order to personalize the online information based on the user's desires.

Techniques used for recommender systems are alternative, user-centric, and promising approaches to undertake the problem of information overload by adapting the content and the structure of the websites and the knowledge obtained from the analysis of the users' access behaviors [4]. Recommender systems satisfy the needs of users without explicit choice made by them.

In general, the recommender systems focus on the process of recognizing web users or objects, accumulating information with respect to users' favorites or interests as well as adapting the services to satisfy the users' needs. Briefly, web recommender systems can be used to provide enhanced quality service of web ap-

*Corresponding author: Rana Forsati, NLP Research Lab, Faculty of Electrical and Computer Engineering, Shahid Beheshti University, GC, Tehran, Iran. E-mail: rana.forsati@gmail.com.

25 plications to users during their browsing period [40–
26 42].

27 Several approaches are introduced for recommender
28 systems, which can be categorized into three main gr-
29 oups, i.e. 1) content-based, 2) collaborative filtering
30 and 3) hybrid systems. [6] has introduced four dif-
31 ferent classes of recommendation techniques based
32 on knowledge sources: 1) collaborative filtering, 2)
33 content-based, 3) knowledge-based and 4) demogra-
34 phic.

35 Collaborative Filtering (CF) is one of the most suc-
36 cessful and extensively used technologies for building
37 recommender systems. The goal of CF is to guess the
38 preferences of a user, known as the active user, based
39 on the preferences of a group of users.

40 Collaborative filtering suffers from a number of
41 well-known disadvantages including the cold start/
42 latency problem, sparseness within the rating matrix,
43 scalability, and efficiency [7].

44 A content-based recommender uses descriptions of
45 the content of the items to find out the relationship
46 between a single user and the description of items.
47 This approach faces several essential defects. It cap-
48 tures only partial information on item characteristics,
49 usually textual information. Other content information
50 such as audio or visual content is usually disregarded.
51 It tends to recommend only items with similar char-
52 acteristics (also known as the over-specification prob-
53 lem). Only the target user's feedback is used in this ap-
54 proach, disregarding the fact that user's interests may
55 also be influenced by other users' interests [8].

56 Hybrid recommender systems combine two or more
57 of the techniques to improve their performance. There
58 are some strategies for hybrid recommendation such as
59 weighted, switching, mixed, feature combination, fea-
60 ture augmentation, cascade, and meta-level [6].

61 Recently, there has been an increasing attention
62 in applying Web mining techniques to building Web
63 recommender systems. It was first proposed by Et-
64 zioni [9]. Web mining is the use of data mining tech-
65 niques to automatically find out and extract informa-
66 tion from Web services and documents [9].

67 The most known classification in Web mining clas-
68 sifies it into three groups: 1) Web content mining,
69 2) Web structure mining, and 3) Web Usage min-
70 ing [4]. Web content mining focuses on the discovery/
71 retrieval of the functional information from the web
72 contents/documents, while the Web structure mining
73 emphasizes how to model the underlying link struc-
74 tures of the Web. Web usage mining describes the tech-
75 niques which discover the user's usage patterns and
76 makes an attempt to predict the user's behaviors [10].

Lately, the systems which use merits of a combi-
nation of content, usage and even structural informa-
tion of the websites have been introduced [11–14],
which show superior results on web page recommen-
dation [4].

In [1,2], all three aspects of web mining have been
used to produce recommendations.

A bipartite graph introduced in [20] consists of users
and movies, where each directed edge corresponds to
the user rating the movie. Then, the given task can be
further formulated as a link existence prediction prob-
lem. The key idea in this approach is to simultaneously
obtain user and movie neighborhoods via co-clustering
and then generate predictions based on the results of
co-clustering.

Huang and her colleagues [23] also proposed a two-
layer graph-based recommender system for a digital li-
brary. In this paper, the customer similarity has been
calculated via the demographic information of cus-
tomers, and the similarity of books has been com-
puted by using content and attribute information of the
books.

Forsati et al. [18] proposed an algorithm that takes
advantage of usage data and link information to rec-
ommend pages to users. The algorithm is based on
distributed learning automata and the PageRank algo-
rithm.

Mohammadi Doustdar and colleagues [21] intro-
duced a Hybrid recommender system which combines
content-based and collaborative approaches in a bi-
section graph model.

In this paper, in order to improve the recommender
performance, we present a practical and efficient ap-
proach which is a combination of content-based and
collaborative filtering approaches in a two-layer graph
model getting use of web content and usage mining.

The rest of this paper is organized as follows. In
Sections 2, 3 and 4 we overview the fuzzy C-means
clustering algorithm, genetic k-means approach and an
improved genetic k-means algorithm. We present our
proposed approach in Section 5. Section 6 gives the
performance evaluation of the proposed algorithms in
comparison to association rule based method [24] and
navigation graph [19]. Section 7 concludes the paper.

2. Algorithms-preliminaries

2.1. The fuzzy c-means clustering algorithm

In 1969, Ruspini introduced the first model of clus-
tering with fuzzy techniques [25].

Fuzzy C-Means (FCM) is a method of clustering which allows data to belong to two or more clusters. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^N \sum_{j=1}^C u_{ij}^m \|x_i - c_j\|^2 \quad (1)$$

Where $m \in [1, \infty)$ is the fuzzy parameter and it is usually equal to 2. u_{ij} is the degree of membership of x_i in the cluster j ; x_i is the i^{th} of d -dimensional data; c_j is the d -dimension center of the j^{th} cluster and $\|\cdot\|$ is the distance of x_i and the cluster center j .

The function J_m cannot be minimized directly so we should use the iterative algorithms. The degree of membership of x_i in the cluster j , u_{ij} , and the c_j cluster center are updated by:

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \quad (2)$$

$$c_j = \frac{\sum_{i=1}^N u_{ij}^m \cdot x_i}{\sum_{i=1}^N u_{ij}^m} \quad (3)$$

This iteration will stop when:

$$\|u_{ij}^{(l)} - u_{ij}^{(l-1)}\| < \varepsilon \quad (4)$$

2.2. The genetic K-means algorithm

Krishna and Murty [26] have combined the k-means clustering and genetic algorithms and have developed the genetic k-means algorithm (GKA). GAs use global search to find optimal solution while K-means use local search. In local search, the obtained solution is usually in the adjacency of previous stage solution. The update of cluster centers in K-means algorithm shows that this algorithm only searches a limited area in the adjacency of initial cluster centers in order to find the optimal solution, while a GA searches a broad area (i.e. globally).

In spite of its high speed, this algorithm is sensitive to initial cluster centers, which increases the probability of selecting local optimal points and affects the final solution.

So the strengths and weaknesses of k-means and genetic algorithms are complementary of each other. Genetic algorithm does well to find the area of research space that probably contains a solution, it is unable

to accurately find the actual location. However, the K-means algorithm does well to find the actual location but it needs the overall view.

So, using a combination of both of these algorithms seems to be a better idea than using only one of them.

An improved genetic K-means algorithm, IGKM, has been introduced in [27]. This algorithm has two outstanding characters. The number of clusters and the fitness function.

Steps of this algorithm are population initialization, clustering, fitness computation, genetic operators, and stopping criteria.

In fitness computation the Inner-cluster distance¹ is defined as:

$$E_k = \sum_{j=1}^k \sum_{i \in I_j} \|x_i - c_j\|^2 \quad (5)$$

where k is the number of clusters, I_j indicates the set of indices of patterns assigned to cluster j , c_j is center of cluster j . Inter-cluster distance² is defined as:

$$D_k = \max_{i,j=1}^k \|c_i - c_j\|^2 \quad (6)$$

where c_i, c_j are respectively the center of i cluster and j cluster. Fitness function is defined as:

$$Fitness(k) = \frac{1}{k} \times \frac{E_1}{E_k} \times D_k \quad (7)$$

The details of this algorithm have been discussed in [27].

3. The proposed approach

The proposed algorithm uses the web content and web usage data based on two-layer graph approach to recommend web pages to the current user. Also, this algorithm uses combination of content-based and collaborative filtering approaches for better performance. For this purpose, we propose a novel fuzzy clustering algorithm based on genetic algorithm and create two individual layers: the layer of web pages, and the layer of users. Then, we apply this fuzzy clustering algorithm on both layers of the graph. Afterwards, According to the carried out clustering, we obtain the similarity between web pages and between users and propose the

¹IND.

²ITD.

recommendations by all the three approaches: content, collaborative, and hybrid approaches.

Due to different membership degrees of data in clusters, the fuzzy clustering algorithms enjoy remarkable performance compared to the hard clustering algorithms.

Also, using of the hybrid approach overcomes the weaknesses of the content-based and collaborative approaches and increases the performance of the recommender system.

Now, we discuss details of the proposed algorithm.

3.1. The fuzzy improved genetic c-means algorithm

Fuzzy Improved Genetic C-means Algorithm, henceforth is called FIGCM, has two outstanding characters similar to IGKM.

The string of chromosome is represented by the number of clusters k . We use the coding of cluster centers and suppose that the number of clusters in the first population is $k \leq \sqrt{n}$.

Stages of this algorithm are as follow:

1. Population initialization. The population P is generated randomly.
2. Clustering.
Determining the membership degree of each data in each cluster. The membership degree of each data in each cluster is computed as follow:

$$u_{ij} = \frac{\left(1 / \|x_i - c_j\|^2\right)^{1/m-1}}{\sum_{k=1}^C \left(1 / \|x_i - c_k\|^2\right)^{1/m-1}} \quad (8)$$

for $i = 1, \dots, N$ and $j = 1, \dots, C$

Where m is the fuzzy parameter and is usually equal 2. u_{ij} is the membership degree of x_i in j^{th} cluster, c_j is the center of j^{th} cluster, and $\|\cdot\|$ is the distance of data to the cluster center.

3. Specifying the new cluster centroid based on the membership degree of each data in each cluster, with the following formula:

$$c_j = \frac{\sum_{i=1}^N u_{ij}^m \cdot \vec{x}_i}{\sum_{i=1}^N u_{ij}^m} \quad (9)$$

for $j = 1, \dots, C$

4. Selection operator. We use the roulette wheel selection in this algorithm.
5. Crossover. Suppose the random operator $r \in [0, 1]$, then $r \times A.NumberofCluster$ clusters from chromosome A is combined to $r' \times B.NumberofCluster$ clusters from chromosome B , and the first child is created. The combination of the rest of the cluster centers creates the second child. In this algorithm, the rate of crossover is assumed to be 0.6.
6. Mutation. One of the cluster centers is created randomly, and is added to the random real variable. The rate of mutation in this algorithm is supposed to be 0.3.
7. KMO. The KMO operator [26] is one of the techniques used for faster convergence of GA. The KMO operator in this algorithm is 2.
8. Fitness Evaluations. The fitness function is defined similar to IGKM as follow:

$$E_k = \sum_{i=1}^k \sum_{j=1}^n u_{ij}^m \|x_i - c_j\|^2 \quad (10)$$

where k is the number of clusters, u_{ij} is membership degree of x_i in j^{th} cluster, c_j is the center of the j^{th} cluster, and $\|\cdot\|$ is the distance of data to the cluster center.

ITD is defined:

$$D_k = \max_{i,j=1}^k \|c_i - c_j\|^2 \quad (11)$$

where c_i, c_j are respectively cluster center i and cluster center j . Fitness function is:

$$Fitness(k) = \frac{1}{k} \times \frac{E_1}{E_k} \times D_k \quad (12)$$

3.2. A bi-section graph approach

In this paper, we construct a two-layer graph-based recommender system to combine the content-based and the collaborative filtering approaches (Fig. 1). This graph consists of user and web page layers and incorporates user-to-user correlation, webpage-to-webpage correlation and user-to-webpage correlation. Each node in the web page layer shows a web page and each node in the user layer represents a user [1].

The method we used to construct a two-layer graph consists of the following computational stages:

1. Creating web page layer. Representing web pages by vector of keywords and creating the first layer of the graph.

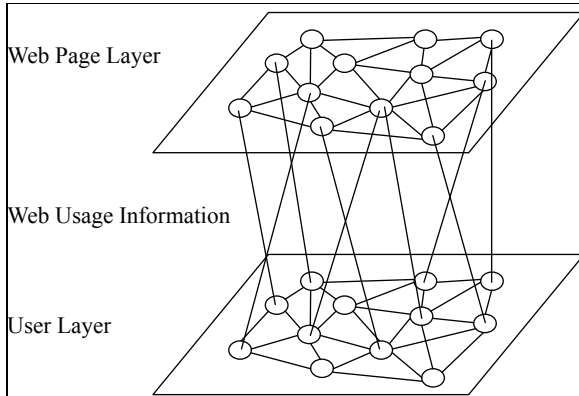


Fig. 1. A bi-section graph model of web pages and users.

- 266 2. FIGCM on web pages. Applying the FIGCM algorithm in the web page layer and obtaining the similarity of web pages.
- 267
- 268
- 269 3. Identification of the sessions. Recognizing sessions in the log file using maximum time of 30 minutes and considering a default threshold for similarity of consecutive web pages [30].
- 270
- 271
- 272
- 273 4. Creating user layer. Finding the weight of each web page in each session, demonstrating sessions by vector of web pages and constructing the second layer of the graph.
- 274
- 275
- 276
- 277 5. FIGCM on users. Applying the FIGCM algorithm in the user layer and obtaining the similarity of users.
- 278
- 279
- 280 6. Constructing correlation between two layers. Creating correlation between these layers based on the calculated weights in stage 4.
- 281
- 282
- 283 7. Content based Recommendation. Proposing the web pages based on content-based approach (in the layer of web pages).
- 284
- 285
- 286 8. Collaborative based Recommendation. Recommending the web pages based on collaborative filtering approach (in the layer of users).
- 287
- 288
- 289 9. Hybrid Recommendation. Suggesting recommendation by the hybrid approach (combination of content-based and collaborative approaches).

292 A summary of the recommendation process is shown in Fig. 2. This model is flexible, comprehensive, and modular [8].

295 Firstly, it is flexible because we can control the parameters easily without building a new model. Secondly, this model includes three approaches of recommendation, content-based, collaborative and hybrid approaches, which can be applied in the comprehensive model. Thirdly, this model is modular and allows for future expansion. Since two layers of graph are inde-

Recommendation Process:

1. Create Graph Layer (1, Web Pages)
2. FIGCM (Graph Layer1)
3. Session Identification
4. Create Graph Layer (2, Users)
5. FIGCM (Graph Layer2)
6. Calculate Correlation Weights
7. Content based Recommendation
8. Collaborative based Recommendation
9. Hybrid Recommendation

Fig. 2. A summary on recommendation process.

pendent of each other, we can adopt different algorithmic techniques on each stage to test different performances. For example, we can change the clustering algorithm in web page layer without changing the method used for users' clustering.

3.2.1. Web page representation

In the vector-space model, each web page is represented by the vector of weights of n keywords as follows:

$$d_i = (w_{i1}, w_{i2}, \dots, w_{ij}, \dots, w_{in}) \quad (13)$$

where w_{ij} is the weight of keyword j in web page i and n is the total number of the unique keywords. The most widely used weighting schema is the combination of term frequency and inverse document frequency (TF-IDF) [31,32], The TF-IDF score of each w_{ik} can be computed by the following formula [33]:

$$w_{ij} = tf(i, j) \times idf(i, j) = tf(i, j) \times \left(\log \frac{N}{df(j)} \right) \quad (14)$$

where $tf(i, j)$ is the number of occurrences of keyword j in a web page d_i , N is the number of web pages in the whole collection, and $df(j)$ is the number of web pages in which keyword j appears.

3.2.2. User representation

We suppose each session as the sequence of visited web pages. In other words, each session is a vector of web pages' weights. In order to identify sessions in log, we use maximum time of 30 minutes and a similarity threshold between two pages visited consecutively [30].

In order to improve the quality of our recommender system, we have used the importance of the web pages in the sessions. Generally, all of the web pages accessed by a user don't interest him/her with the same rate. Therefore, it is not efficient to use all of the visited pages equally to make recommendation. So we

try to approximate the degree of importance of each web page for users. We signify each session as an m -dimensional vector over the space of web pages, $s \leq (p_1, w_1), (p_2, w_2), \dots, (p_m, w_m) >$, where w_i indicates the i^{th} web page weight ($1 \leq i \leq m$) visited in a sessions [1,43].

For computing the web page weight, we use duration and frequency parameters.

Duration reflects the relative importance of each page, because a user generally spends more time on a more useful page, else if a user is not interested in a page, he/she would not spend much time on that page and usually jumps to another page quickly. However, a quick jump might also occur due to the short length of a web page and the size of a page may affect the actual visiting time. Hence, it is more appropriate to accordingly normalize duration by the length of the web page, that is, the total bytes of the page. Frequency is the number of times a page is accessed by different users. Details of this parameters are discussed in [3,4].

3.2.3. The inter layer links between web page layer and user layer

After creating web page and user layer, we achieve inter-layer correlations computed by web page weight in a session in the previous section. The inter layer links between user layer and web page layer is simply derived from the weight of the web pages in the sessions.

3.3. Applying fuzzy improved genetic c-means algorithm in two-layer graph

In this part, the FIGCM algorithm is applied to web page layer and user layer, and then we obtain the vector $D_i = (d_{i1}, d_{i2}, \dots, d_{ic})$ where each element is the membership degree of that web page in each cluster. If two web pages are more similar, their degrees of membership in different clusters will be closer to each other. Therefore, we compute the similarity of web pages using Euclidean distance of these vectors. Also, we obtain vector $U_j = (d'_{j1}, d'_{j2}, \dots, d'_{jc})$ for each user where each element is the membership degree of that user in each cluster. If two users are more similar, the degrees of their membership to different clusters will be closer to each other. Therefore, we compute the similarity of users using Euclidean distance of these vectors. The resulted clustering is used in the recommendation process.

3.4. Recommendation mechanism

3.4.1. Using FIGCM algorithm in web page layer for content-based recommendations

The web pages that are similar to the visited web pages of the target user are retrieved as content-based recommendations.

3.4.2. Using FIGCM algorithm in user layer for collaborative filtering recommendations

First, a list of users, similar to the target user is obtained. Then, the web pages marked as interested by those users are retrieved as the collaborative filtering recommendations for the target user.

3.4.3. The hybrid recommendations

The hybrid recommendations are obtained by combining the recommendation results from the two approaches described above through the switching strategy.

4. Experimental evaluation

4.1. Data set

In this section, we present a set of experiments that are carried out for evaluating the impact of our proposed techniques on the recommendation process.

We have done preliminary experiments on the Music Machines³ data set [19,34–37]. This web site is collected in September and October of 1997 and is used mainly for experimental purposes.

We have used Music Machines data sets because numerous approaches such as [19,34–37] have applied this data sets and unlike most web traces, this was specifically configured to prevent caching, so the log represents all requests (not just the browser cache misses).

This web site contains information about various kinds of electronic musical instrument grouped by manufacturers [34]. For each manufacturer, there may be multiple entries for different instrument models available – keyboards, electric guitars, amplifiers, etc.

Each access log consists of the user label, request method, accessed URL, data transmission protocol, access time and browser used to access the site. The server logs were filtered to remove those entries that

³<http://www.hyperreal.org/music/machines/>.

are irrelevant for analysis and those referring to pages that do not exist in the available site copy [1].

The data is based on a 2-week log file during 12–25 February of 1997 and the filtered data contains 16427 sessions and 892 web pages. The total number of unique words is 9299.

For this experiment, we divide the resulting set of transactions into a training (approx. 80%) and a testing set (approx. 20%).

4.2. Evaluation methodology and metrics

In order to evaluate our recommender system, we measured the performance of proposed method using two different standard measures, namely Precision and Coverage [38]. Recommendation precision and coverage are two metrics quite similar to the precision and recall metrics that are usually used in information retrieval literature. Recommendation precision measures the ratio of correct recommendations (i.e., the proportion of relevant recommendations to the total number of recommendations), where correct recommendations are the ones that appear in the remaining section of the user session. For each visit session after considering each page p , the system generates a set of recommendations $R(p)$. To compute the Precision, $R(p)$ is compared with the rest of the session $T(p)$ as follows:

$$\text{Precision} = \frac{T(p) \cap R(p)}{R(p)} \quad (15)$$

On the other hand, recommendation coverage shows the ratio of the pages in the user session that the system is able to predict) i.e., the proportion of relevant recommendations to all pages that should be recommended) before the user visits them:

$$\text{Coverage} = \frac{T(p) \cap R(p)}{T(p)} \quad (16)$$

To find an optimal trade-off between precision and coverage a measure like the E-measure [39] can be used. The parameter manages the trade-off between precision and coverage.

$$\text{E-measure} = \frac{1}{\alpha(1/\text{Precision}) + (1 - \alpha)(1/\text{Coverage})} \quad (17)$$

A popular single-valued measure is the F-measure. It is defined as the harmonic mean of precision and coverage.

$$F - \text{Measure} = \frac{2 * \text{Precision} * \text{Coverage}}{\text{Precision} + \text{Coverage}} \quad (18)$$

It is a special case of the E-measure with $\alpha = 0.5$.

4.3. Results and discussions

We evaluate our method under different settings. The first experiments were performed to evaluate system sensitivity to the size of visit window ($|w|$, the portion of user histories used to produce recommendations) and recommendation window ($|w'|$). We show the effect of them on efficiency of the proposed system.

4.3.1. Impact of active window size on user navigation trail

In all experiments, we measured F-Measure of recommendations against varying number of recommended pages. In our state definition, we used the notion of N-Grams by putting a sliding window on user navigation path. The implication of using a sliding window of size w is that we base the prediction of user future visits on his w past visits. The choice of this sliding window size can affect the system in several ways. To consider the impact of window size on the FIGCM algorithm, we also vary window sizes from 1 to 12.

The impact of different window sizes on F-Measure scores of recommendations against varying the number of recommended pages from 1 to 15 is depicted in Figs 3–5. These figures demonstrate the F-Measure of our proposed approaches, i.e. content-based approach, collaborative approach and hybrid approach, respectively. A large sliding window seems to provide more information to the system while on the other hand causing a larger state space with sequences that occur less frequently in the usage logs. We evaluated our performance system with different window sizes on user trail as seen in these Figures.

As our experiments show, the best results in the content-based approach are achieved when window size = 3 and recommendation window = 2, in the collaborative approach the best results appear when window size = 4 and recommendation window = 2, and in the hybrid approach it seems better to set the window size = 4 and recommendation window = 2.

In all of these three approaches, the maximum of F-Measure belongs to the recommendation window 2 and 3.

It can be inferred from this diagrams that a window of size 1 ($|w| = 1$) which considers only the user's last page visit does not hold enough information to make the recommendation. The F-Measure of recommendations improves with increasing the window size and the best results are achieved with a window size of 3, 4 in difference approaches. As shown in Figs 3–5 using a window size larger than 4 decreases the system perfor-

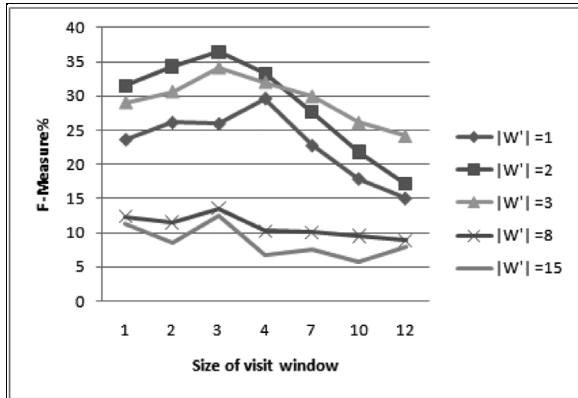


Fig. 3. F-Measure in content-based approach for various size of visit window and recommendation window. (Colours are visible in the online version of the article; <http://dx.doi.org/10.3233/HIS-130166>)

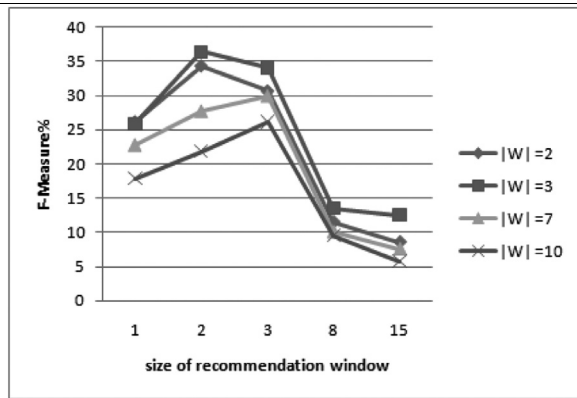


Fig. 6. F-Measure in content-based approach for various size of recommendation window and visit window. (Colours are visible in the online version of the article; <http://dx.doi.org/10.3233/HIS-130166>)

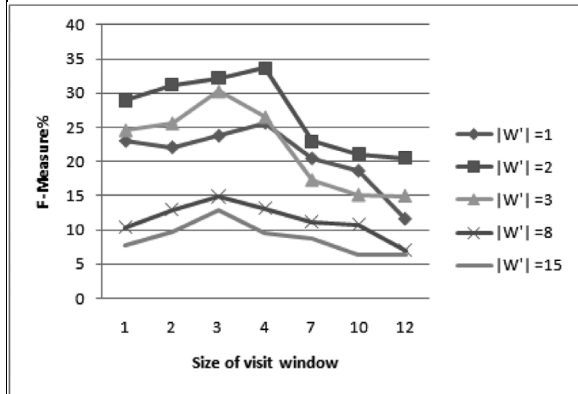


Fig. 4. F-Measure in collaborative approach for various size of visit window and recommendation window. (Colours are visible in the online version of the article; <http://dx.doi.org/10.3233/HIS-130166>)

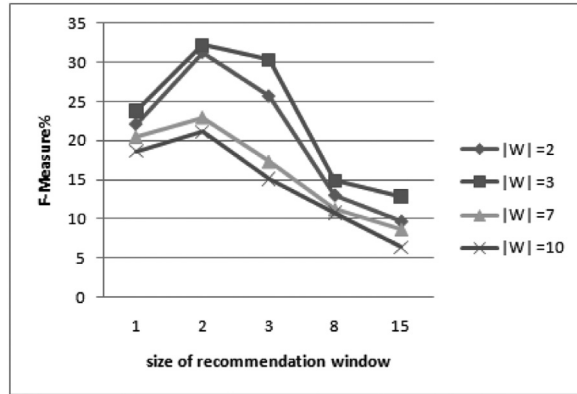


Fig. 7. F-Measure in collaborative approach for various size of recommendation window and visit window. (Colours are visible in the online version of the article; <http://dx.doi.org/10.3233/HIS-130166>)

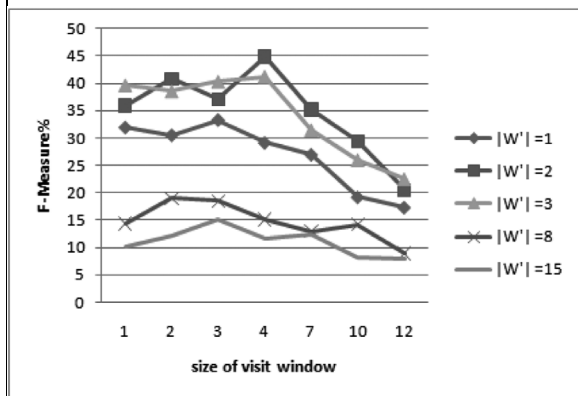


Fig. 5. F-Measure in hybrid approach for various size of visit window and recommendation window. (Colours are visible in the online version of the article; <http://dx.doi.org/10.3233/HIS-130166>)

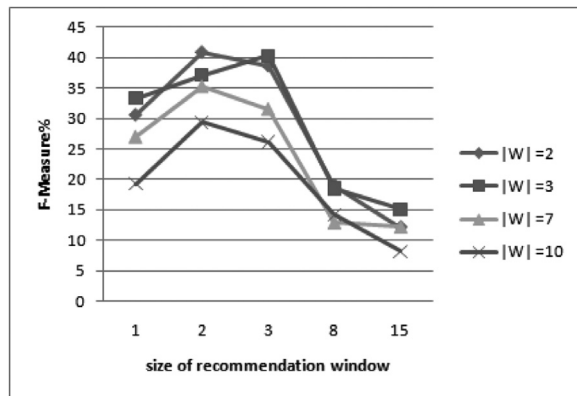


Fig. 8. F-Measure in hybrid approach for various size of recommendation window and visit window. (Colours are visible in the online version of the article; <http://dx.doi.org/10.3233/HIS-130166>)

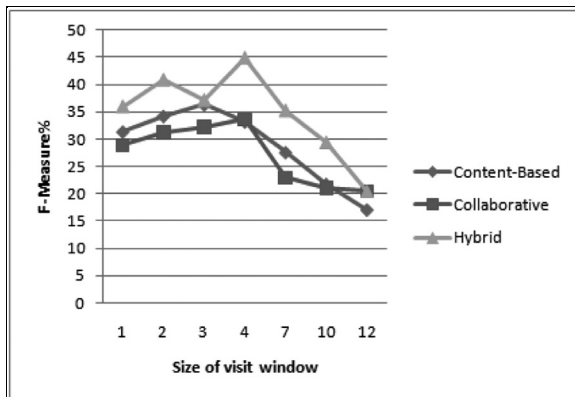


Fig. 9. Comparing F-Measure in content-based, collaborative, hybrid approach. (Colours are visible in the online version of the article; <http://dx.doi.org/10.3233/HIS-130166>)

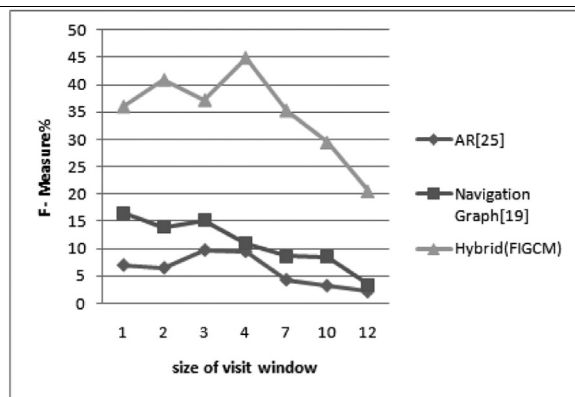


Fig. 10. Comparing F-Measure in hybrid, AR and navigation graph approach. (Colours are visible in the online version of the article; <http://dx.doi.org/10.3233/HIS-130166>)

508 mance, this means the large value of w leads in unde-
509 sirable recommendations.

510 The impact of different recommendation window
511 sizes on F-Measure against various visit window sizes
512 from 2 to 10 is depicted in Figs 6–8.

513 The results show that in recommendation window
514 larger than 3, the F-Measure decreases. The best per-
515 formance is in the hybrid approach in window size 4
516 and recommendation window 2.

517 4.3.2. Comparison with other methods

518 As our experiments on the previous section show the
519 best performance is in the hybrid approach in window
520 size 4 and recommendation window 2.

521 On the other hand, the mean of transaction length is
522 5; in these experiments we have used a fixed recom-
523 mendation window 2 and different values for window
524 size.

525 We first compared three recommender systems,
526 Content-based Recommender algorithm, Collaborative
527 filtering Recommender algorithm and Hybrid algo-
528 rithm to each other. The Recommendation F-Measure
529 of the three systems is depicted in Fig. 9.

530 Comparison of the proposed systems indicates that
531 the hybrid approach gains much better results than
532 content-based and collaborative filtering approaches,
533 because the hybrid approach eliminates limitations and
534 weaknesses of each of them.

535 This figure verifies our justification for using two al-
536 gorithms in building a hybrid recommender system.

537 We observe our system performance in compari-
538 son with Association Rules (AR), which is commonly
539 known as one of the most successful approaches in
540 web mining based recommender systems [24] and a

541 graph based recommender system discussed in [19].
542 Figure 10 shows the comparison of hybrid system's
543 performance with AR method and navigation graph ap-
544 proach in the sense of their F-Measure in recommen-
545 dation window = 2 and difference window sizes on
546 Music Machines dataset.

547 Experimental results show that our hybrid approach
548 improves performance significantly and gains much
549 better results than AR method and navigation graph.

550 The hybrid approach has been determined to be ca-
551 pable of making web recommendation more accurate
552 and effective than the conventional methods. In sum-
553 mary, this experiment shows that our system can sig-
554 nificantly improve the quality of web site recommen-
555 dation by combining two information channels, while
556 each channel includes contributions to this improve-
557 ment.

558 5. Conclusions and future work

559 In this paper we proposed a new hybrid method for
560 web page recommendation. First, we produced the rec-
561 ommendations based on content-based approach in the
562 first graph layer and, then, based on collaborative filter-
563 ing approach in the second graph layer. By introducing
564 the hybrid algorithm, we present our third algorithm in
565 which the switching method is used for the combina-
566 tion of the two above approaches.

567 In the proposed approach, the web pages in a rec-
568 ommendation list are ranked according to their impor-
569 tance, which is in turn computed based on web content
570 and usage information.

571 One of the challenging problems in recommenda-
572 tion systems is dealing with unvisited or newly added

pages. This problem is solved with the novel hybrid approaches.

Our experimental results illustrate that using this hybrid algorithm in a web recommender system has the potential to improve the quality of the system and it can generate higher quality recommendations than using either the content-based recommendation or the collaborative filtering recommendation algorithm alone. The results show that the proposed model can significantly improve the recommendation effectiveness.

References

- [1] H.M. Doustdar, R. Forsati, M.R. Meybodi and M. Shamsfard, A bi-section graph approach for hybrid recommender system, in: *GRC, 2011 IEEE International Conference on Granular Computing*, 2011.
- [2] H.M. Doustdar, R. Forsati and M.R. Meybodi, The hybrid web recommender system based on two-layer graph and graph partitioning, in: *Proceeding of the 5th Data Mining Conference*, Tehran, Iran, 2011.
- [3] M. Talabeigi, R. Forsati and M.R. Meybodi, A hybrid web recommender system based on cellular learning automata, in: *GRC, 2010 IEEE International Conference on Granular Computing*, 2010, pp. 453–458.
- [4] R. Forsati and M.R. Meybodi, Effective web page recommendation algorithms based on distributed learning automata and weighted association rules, *Journal of Expert Systems with Applications*, 2010, 1316–1330.
- [5] R. Forsati, M.R. Meybodi and A. Ghari Neiat, Web page personalization based on weighted association rules, in: *The International Conference on Electronic Computer Technology*, 2009, pp. 130–135.
- [6] R. Bruke, *Hybrid Recommender Systems*, School of Computer Science, Telecommunications and Information Systems, Springer Berlin Heidelberg, 2007, pp. 377–408.
- [7] M.O. Mahony, N. Hurley, N. Kushmerick and G. Silverstre, Collaborative recommendations: A robustness analysis, *ACM Trans, Internet Tech* 4(4) (2004), 344–377.
- [8] Z. Huang, W. Chung and H. Chen, A graph model for e-commerce recommender systems, *Journal of the American society for information science and technology*, (2004), 259–274.
- [9] O. Etzioni, The world wide web: Quagmire or gold mine, *Communications of the ACM* 39(11) (1996), 65–68.
- [10] Y. Wang, *Web Mining and Knowledge Discovery of Usage Patterns*, 2000.
- [11] M. Eirinaki, M. Vazirgiannis and I. Varlamis, SEWeP: Using site semantics and taxonomy to enhance the web personalization process, in: *Proceeding of the 9th SIGKDD Conference*, 2003.
- [12] M. Eirinaki, C. Lamos, S. Paulakis and M. Vazirgiannis, Web personalization integrating content semantics and navigational patterns, in: *Proceedings of the sixth ACM workshop on Web Information and Data Management WIDM*, 2004.
- [13] J. Li and O.R. Zaiane, Combining usage, content and structure data to improve web site recommendation, in: *5th International Conference on Electronic Commerce and Web*, 2004.
- [14] B. Mobasher, H. Dai, T. Luo, Y. Sun and J. Zhu, Integrating web usage and content mining for more effective personalization, in: *EC-Web*, 2000, pp. 165–176.
- [15] M. Nakagawa and B. Mobasher, A hybrid web personalization model based on site connectivity, in: *The Fifth International WEBKDD Workshop: Web mining as a Premise to Effective and Intelligent Web Applications*, 2003, pp. 59–70.
- [16] B. Mobasher, Web Usage Mining and Personalization, in: *Practical Handbook of Internet Computing*, M.P. Singh, ed., CRC Press, 2005.
- [17] B. Mobasher, H. Dai, T. Luo and M. Nakagawa, Improving the effectiveness of collaborative filtering on anonymous web usage data, in: *Proceedings of the IJCAI 2001 Workshop on Intelligent Techniques for Web Personalization (ITWP01)*, 2001.
- [18] R. Forsati, M.R. Meybodi and M. Mahdavi, Web page personalization based on distributed learning automata, in: *Proceedings of the Third Information and Knowledge Technology*, Ferdowsi University of Mashad, Mashad, Iran, Nov. 27–29, 2007.
- [19] Y. Wang, W. Dai and Y. Yuan, Website browsing aid: A navigation graph-based recommendation system, *Journal Decision Support systems* (2008).
- [20] T. Liu, Y. Tian and W. Gao, A two-phase spectral bigraph co-clustering approach, in: *KDD Cup and Workshop 2007, at the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007.
- [21] H. Mohammadi Doustdar, R. Forsati, M.R. Meybodi and M. Shamsfard, A bi-section graph approach for hybrid recommender system, in: *Proceedings of IEEE International Conference on Granular Computing*, 2011, pp. 171–176.
- [22] V.A. Koutsonikola and A. Vakali, A fuzzy bi-clustering approach to correlate web users and pages, *International Journal of Knowledge and Web Intelligence* (2009), 3–23.
- [23] Z. Huang, W. Chung, T.H. Ong and H. Chen, A graph-based recommender system for digital library, in: *ACM/IEEE Joint Conference on Digital Libraries*, 2002, pp. 65–73.
- [24] B. Mobasher, H. Dai, T. Luo and M. Nakagawa, Effective personalization based on association rule discovery from web usage data, in: *Proceedings of the 3rd ACM Workshop on Web Information and Data Management (WIDM01)*, Atlanta, Georgia, 2001.
- [25] A. Baraldi and P. Blonda, A survey of fuzzy clustering algorithms for pattern recognition – Part I and II, *IEEE Transactions on Systems, Man, and Cybernetics* 29(6) (1999).
- [26] K. Krishna and M.N. Murty, Genetic k-means algorithm, *IEEE Transactions on Systems, Man, and Cybernetics* 29(3) (1999), 433–439.
- [27] H.X. Guo, K.J. Zhu, S.W. Gao and T. Liu, An improved genetic k-means algorithm for optimal clustering, in: *Sixth IEEE International Conference on Data Mining-Workshops (ICDMW'06)*, 2006.
- [28] Y. Lu, S. Lu, F. Fotouhi, Y. Deng and S.J. Brown, Incremental genetic K-means algorithm and its application in gene expression data analysis, *BMC Bioinformatics* 5(172) (2004).
- [29] C.A. Murthy and N. Chowdhury, In search of optimal clusters using genetic algorithms, *Pattern Recog Lett* (1996), 825–832.
- [30] J. Li and O.R. Zaiane, Combining usage, content and structure data to improve web site recommendation, in: *5th International Conference on Electronic Commerce and Web*, 2004, pp. 305–315.
- [31] B. Everitt, *Cluster Analysis*, (2nd Edition), Halsted Press, New York, 1980.
- [32] G. Salton, *Automatic text processing*, Addison-Wesley, 1989.
- [33] G. Salton and C. Buckley, Term-weighting approaches in automatic text retrieval, *Information Processing and Manage*

- 696 *ment: an International Journal* **24**(5) (1988), 513–523. 713
- 697 [34] M. Perkowitz and O. Etzioni, Adaptive web sites: Automati- 714
698 cally synthesizing web pages, in: *Proceedings of The Fifteenth* 715
699 *National Conference On Artificial Intelligence*, 1998. 716
- 700 [35] N. Kushmerick, J. McKee and F. Toolan, Towards zero-input 717
701 personalization: Referrer-based page prediction, **1892** (2000), 718
702 133–143. 719
- 703 [36] M. Perkowitz and O. Etzioni, Towards adaptive web sites: 720
704 Conceptual framework and case study, *Journal of Artificial* 721
705 *Intelligent* **118** (2000), 245–275. 722
- 706 [37] Y. Chen, X. Chen and H. Chen, Improve on frequent ac- 723
707 cess path algorithm in web page personalied recommendation 724
708 model, in: *International Conference on Information Science* 725
709 *and Technology*, 2011. 726
- 710 [38] C. Ziegler, G. Lausen and L. Schmidt-Thieme, Taxonomy- 727
711 driven computation of product recommendations, in: *Pro-* 728
712 *ceedings of the ACM Conference on Information and Knowl-*
edge Management, 2004, pp. 406–415.
- [39] V. Rijsbergen, *Information Retrieval*, Butterworth, London, 1979.
- [40] J.D. Velásquez and V. Palade, *Adaptive Web Sites: A Knowl-*
edge Extraction from Web Data Approach, IOS Press, 2008.
- [41] J.D. Velasquez and V. Palade, Building a knowledge base
for implementing a web-based computerized recommenda-
tion system, in: *International Journal on Artificial Intelli-*
gence Tools **16**(5) (2007), 793–828.
- [42] J.D. Velásquez and V. Palade, A knowledge base for the main-
tenance of knowledge extracted from web data, *Knowledge-*
Based Systems **20**(3) (2007), 238–248.
- [43] M. Talabeigi, R. Forsati and M.R. Meybodi, A hybrid web
recommender system based on cellular learning automata, in:
2010 IEEE International Conference on Granular Computing
(GrC), Aug 2010, pp. 453–458, doi: 10.1109/GrC.2010.153.