A Bi-section Graph Approach for Hybrid Recommender System

Hanieh Mohammadi Doustdar
Department of Computer Engineering
Islamic Azad University, Qazvin Branch
Qazvin, Iran
h.m.doustdar@gmail.com

Rana Forsati
Electrical & Computer Engineering
Shahid Beheshti University, G. C.,
Tehran, Iran
r_forsati@sbu.ac.ir

Mohammad Reza Meybodi
Department of Computer Engineering,
Amirkabir University of Technology
Tehran, Iran
mmeybodi@aut.ac.ir

Mehrnoush Shamsfard
Electrical & Computer Engineering
Shahid Beheshti University, G. C.,
Tehran, Iran
shams@sbu.ac.ir

Abstract— **With the rapid growth of the World Wide Web, many efforts have been done to address the problem of information overload. Recommender systems help users make decisions in this huge information space. Most existing recommender system use either content-based or collaborative approach. It could be difficult to estimate a single best model for recommendation. Each of the single methods has own strengths, but also limitations and weaknesses. Therefore, combination of different methods can overcome these shortcomings and may result in better accuracy. In this paper, in order to have a better performance, we have introduced a hybrid recommender system which combines theses approaches in a bi-section graph model. We have gained web page similarity and user similarity by the new methods and have modeled web pages and users in the two-layer graph. Evaluation results show that combining these approaches achieved more accurate predictions and relevant recommendations than using only one of them.**

Keywords-recommender systems; content-based approach; collaborative filtering approach; two-layer graph

I. INTRODUCTION

In recent years, the massive influx of information onto World Wide Web has facilitated users, not only by retrieving information, but also discovering knowledge. However, web users usually suffer from the information overload problem due to the fact of significantly increasing and rapidly expanding growth in amount of information on the web. To deal with this problem, researchers have proposed recommender systems which automatically select and recommend web pages suitable for user's favor. Given a user's current actions, the goal is to determine which web pages will be accessed next. Many web sites use web page recommender systems to increase their usability and user satisfaction.

One research area that has recently contributed greatly to this problem is web mining.

Web mining is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World Wide Web. It includes web content data, web structure data and web usage data.

In this paper, we have used all three categories of web mining to propose a two-layer graph based recommender system.

In [1] distinguished four different classes of recommendation techniques based on knowledge source: collaborative filtering, content-based, knowledge-based, Demographic.

Numerous approaches are introduced for personalization system use from collaborative filtering and content-based approach.

Collaborative Filtering is one of the most successful and widely used technologies for building recommendation systems. It makes prediction for a user based on the similarity between the interest profile of that user and those of other users. This approach is used in [2,3].

Collaborative filtering does not work well when an item is newly introduced or a user just starts to use the system, because the system does not have much rating information on either the item or the user (the "cold-start problem"). Also, when there are many items but relatively few users, collaborative filtering cannot provide a good recommendation simply because there are few ratings (the "sparsity problem"). Furthermore, some users with opinions consistently different from the group opinions do not benefit from collaborative filtering (the "gray sheep problem") [13].

The content-based recommender uses descriptions of the content of the items to learn the relationship between a single user and the description of the items. Systems like those in [4,5] are examples using this method.

This approach has several fundamental limitations. It captures only partial information on item characteristics, usually textual information. Other content information such as audio or visual content is usually ignored. This approach tends to recommend only items with similar characteristics (also known as the "over-specification problem"). Only the target user's feedback is used in this approach, although that user's interest may also be influenced by other users' preferences [13].

The performance of a recommender model depends on the structure of the web site besides the specific technique that it uses. Furthermore, different users may have different navigation strategies. Thus, it could be difficult to estimate a single best model for recommendation. Each of the single methods has own strengths, but also limitations and weaknesses.

978-1-4577-0371-3/11/$26.00 ©2011 IEEE

Therefore, combination of different methods can overcome these shortcomings and may result in better accuracy.

Hybrid recommender systems combine two or more of the techniques to improve recommender performance [1,6].

In this paper, in order to improve recommender performance, we have presented an elegant and effective framework for combining content-based and collaborative filtering approaches in a two-layer graph model. This model represents the user-webpage information which contains nodes (users and web page) and links (similarities and weight of web pages in sessions). We use a Music Machines[1] web site as our test-bed to implement a recommender system.

Some recommender systems designed based on graph-theoretic are as follows:

The horting approach, proposed by Aggarwal et al. [7], is a graph collaborative filtering algorithm based on the concepts horting and predictability. Horting is a graph-based technique in which nodes are users, and edges between nodes indicate similarity degree between these. Predictions are produced by walking the graph to nearby nodes and combining the opinions of the nearby users. Based on this graph, recommendations can be generated via a few reasonably short directed paths joining multiple users.

In [8], a navigation graph-based recommendation system is proposed, in which the navigation patterns of previous web site visitors are utilized to provide recommendations for newcomers.

A bipartite graph introduced in [9] consists of users and movies which each directed edge corresponds to the user rated the movie. Then the given task can be further formulated as a link existence prediction problem. 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.

Dhillon [10] was the first to use spectral graph partitioning on a bipartite graph of documents and words, effectively clustering groups of documents and words simultaneously. Consequently, every document cluster has a direct connection to a word cluster; the document clustering implies a word clustering and vice versa.

Koutsonikola et al. [11] computed similarity of users and created a weighted undirected graph, then clustered users. Afterwards, the pages which are mostly visited by the same set of users' clusters are expected to present high value of similarity, while pages visited by different users' clusters present low similarity value, then graph of web pages is constructed and these pages are clustered, too. Thenceforth, users and web pages clusters were assigned to fuzzy logic i.e. groups of related users are interested in a different degree to different groups of related web pages.

Huang et al. [12], also proposed a two-layer graph-based recommender system for digital library. In this paper, the customer similarity has been calculated via demographic information of customers and the similarity of books has been computed by content and attribute information of books. Afterwards, correlations of two layers was obtained by customer transaction information and recommendations was generated based on the association strengths between a customer and the books.

The organization of the paper is as follows: we have introduced a new method for measuring similarity between web pages and users in section III and model users and web pages in the two-layer graph. Section IV discusses about correlation between two layers. Section V describes the recommendation process and section VI presents experimental result. Finally, the paper concludes with future work.

II. A TWO-LAYER GRAPH APPROACH

In this paper, we have introduced a two-layer graph-based recommender system to combine the content-based with the collaborative filtering approach. 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 while each link between any two web pages represents the similarity between them. Each node in the user layer represents a user, and links between user nodes are the similarity between the two users. These similarities will be explained in section III.A, III.B. The inter-layer links are based on weight of web pages in sessions that will be discussed in section IV.

Our method to construct two-layer graph consists of these computational stages:

a) Representing web pages by vector of keywords and computing their content similarities.

b) Finding location of web pages in web site hierarchy and calculating web pages similarity according to this location.

c) Considering weighted mean of two calculated similarities in two previous stages as total web page similarity and creating the first layer of graph.

d) Identifying sessions in the log file using max time=30 minutes and considering predefined threshold for similarity of consecutive web pages.

e) Finding weight of each web page in each session and displaying sessions by vector of web pages and computing session similarity.

f) Calculating session similarity via weighted Levenshtein distance based on similarity of web pages.

g) Considering weighted mean of two computed similarities in two previous stages as total session similarity and constructing second layer of graph.

h) Creating correlation between these layers on account of calculated weight in stage “e”.

We believe that this model is flexible, comprehensive and modular [13].

[1] http://www.hyperreal.org/music/machines/

Firstly, the similarity weights computed in the stages “c”, “g” can be flexibly adjusted to reflect the importance of certain aspects of the similarity. For example, if we want to emphasize the importance of content similarity in our recommendation, we can simply assign a higher weight to it in weighted mean. Finally we believe that this model is flexible because we can control the parameters easily without building new models.

Secondly, this model contained the three approaches of recommendation, content-based, collaborative and hybrid approaches, can be applied in the comprehensive model.

Thirdly, this model is modular and allows for future expansion. Since the computation of similarity between web pages and users and creation of two-layer graph are independent of each other, we can adopt different algorithmic techniques on each stage to test different performances. For example, we can change web page similarity method without changing the method we use for users’ similarity.

III. Evaluating similarity of web sessions

The problem of computing session similarity through a web site is faced by combining the computation of web page similarity and sequence similarity, spent time in each web page, web page size, and visit count of web pages in sessions.

A. *Similarity between Web Pages*

For measuring web page similarity, we use web page content similarity and web page topic similarity.

1) *Web Page Content Similarity:* We represent web pages by using the vector space model. In this model, each web page is represented by a vector of weights of *n* “keywords”: $p_i = (w_{i1}, w_{i2}, \ldots, w_{ij}, \ldots, w_{in})$ where the weight *“w_{ij}”* is the frequency of keyword *j* in web page *i* and *“n”* is number of keywords. The most widely used weighting schema is the combination of Term Frequency and Inverse Document Frequency (TF-IDF) [14], which is defined as

$$w_{ij} = TFIDF(i,j) = tf(i,j).(\log \frac{N}{df(j)}), \quad (1)$$

The content similarity of web pages is

$$Sim_{(Content)}(p_i, p_j) = \frac{\sum_{k=1}^{n} w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^{n} (w_{ik})^2} \sqrt{\sum_{k=1}^{n} (w_{jk})^2}}. \quad (2)$$

2) *Web Page Topic Similarity:* By analyzing the pages, we understand that pages are not randomly spread, but they are organized into a hierarchical structure, called web site hierarchy. A web site hierarchy is a partial order of web pages, in which a leaf node represents a web page corresponding to a file in the server. A non-leaf node in a web site hierarchy represents a web page corresponding to a directory in the server. We use method discussed in [15] with reference to this hierarchy structure called web page topic similarity.

Now, we consider weighted mean of web page content and topic similarity as web page Content Topic similarity.

$$Sim_{(Content\_Topic)}(P_i, P_j) = \alpha * Sim_{(Content)}(P_i, P_j) + \beta * Sim_{(Topic)}(P_i, P_j). \quad (3)$$

Also, in according to links connecting web pages, we use the shortest link between those for measuring web page similarity. Therefore, the formula for calculating web page total similarity can be expressed as follows:

$$Sim_{(Total)}(P_i, P_j) = \frac{1}{Min\_LinkLenght\ (P_i, P_j)} * Sim_{(Content\_Topic)}(p_i, p_j). \quad (4)$$

B. *Similarity between Web User Sessions*

Our basic idea of measuring session similarity is to consider each session as sequence of web page visits.
First we consider each session as vector of web pages and compute web page weight in it and compute session similarity via cosine similarity, afterwards use weighted Levenshtein distance to find the best matching between two sessions. We use from web page similarity which discussed in previous section as one of the sessions similarity metrics in weighted Levenshtein distance. For identifying sessions in log, we use max time=30 minutes and similarity of web pages for two pages visited consecutively. If the similarity is more than a predefined threshold, we will consider it a new session.

1) *Session Similarity based on Weight of Web Pages in Session:* To improve the quality of our recommender system, we use the importance of web pages in the sessions. In general, all accessed pages can be considered interesting to various degrees because users visited them. It is quite probable that not all the pages accessed by the user are of interest to him. Therefore, it is not efficient to use all the visited pages equally to make recommendation. So we try to approximate the degree of importance and interest of a web page for users. we represent each session as an m-dimensional vector over the space of web pages, *$s = <(p_1,w_1),(p_2,w_2),(p_m,w_m)>$*, where w_i denotes the ith web page weight ($1 \le i \le m$) visited in a session s. The web page weight have computed by the method discussed in [16,20].

Therefore, the session similarity is expressed as

$$Sim_{(WebPageWeightInSession)}(S_\alpha, S_\beta) = \frac{\sum_{k=1}^{m} w_{\alpha k} w_{\beta k}}{\sqrt{\sum_{k=1}^{m} (w_{\alpha k})^2} \sqrt{\sum_{k=1}^{m} (w_{\beta k})^2}}. \quad (5)$$

2) *Session Similarity based on Weighted Levenshtein Distance:* We will assume that *A* is a string of length *p* and *B* is a string of length *q*, and that *p≤q*. The edit distance *(A, B)* between strings *A* and *B* is defined as the minimum number of edit operations needed in converting *A* into *B* or vice versa. The Levenshtein edit distance [17] allows three edit operations, which are inserting, deleting or substituting

a character. In this paper we use weighted Levenshtein distance (WLD) [18] which is a slight generalization of the Levenshtein distance (LD). We consider the sessions to be the sequence of web pages, If S_α can be transformed into S_β by substitution of k_i symbols, the insertion of m_i symbols, and the deletion of n_i symbols, then the weighted Levenshtein distance from S_α to S_β is defined as

$$WLD(S_\alpha, S_\beta) = \min_i (p.k_i + q.n_i + r.m_i). \tag{6}$$

where in our proposed method, we consider the $q=r=1$ and $p=1\text{-}sim_{(Total)}(p_i,p_j)$. In other words, we apply page similarity in weighted Levenshtein algorithm to find the best matching between two sessions based on their similarity of pages. Therefore, session similarity can be represented as follows:

$$Sim_{(Levenshtei\ n)}(S_\alpha, S_\beta) = 1 - \frac{Levenshtei\ n(S_\alpha, S_\beta)}{Max(\| E(S_\alpha) \|, \| E(S_\beta) \|)}. \tag{7}$$

Now, we consider weighted mean of session similarity in sections III.B.1 and III.B.2 as session total similarity.

$$\begin{aligned} Sim_{(Total)}(S_i, S_j) &= \alpha * Sim_{(Levenshtei\ n)}(S_i, S_j) \\ &+ \beta * Sim_{(WebPageWei\ ghtInSessi\ on)}(S_i, S_j). \end{aligned} \tag{8}$$

IV. The Inter Layer Links between web page layer and user layer

After calculating page similarity and session similarity, we obtain inter-layer correlation computed by the "equation (5)" which computes the web page weight in a session. The inter layer links between session layer and web page layer is simply derived from weight of web pages in sessions. Each link in the graph has a weight between 0 and 1.

V. Recommendation Process

With the two-layer graph model, all three major recommendation approaches can be implemented by choosing different types of links to use in a recommendation generation process. If only web page information is used, which means only links in the web page layer are activated, it is a content-based approach. If user-layer and inter-layer links are activated, it is a collaborative approach. If all links are activated, it becomes a hybrid approach. We use direct retrieval method to generate recommendations by retrieving web pages similar to the target user's previous visited web pages and web pages visited by users similar to the target user. For content-based recommendation, web pages that are similar to the target user's previous visited web pages are retrieved as recommendations. For collaborative recommendation, a list of users similar to the target user is first obtained. Then, the web pages interested in users are retrieved as the collaborative recommendation for the target user. The hybrid recommendation is obtained by combining the recommendation results from two approaches describe above.

VI. Experimental Results

A. Data Set

In order to evaluate the effectiveness of our proposed method, we have conducted preliminary experiments on Music Machines data sets. This web site is collected in September and October of 1997 and was constructed by professors at Washington University and is used mainly for experimental purposes. We have full access to all documents and access logs.

The Music Machines web site contains information about various kinds of electronic musical equipment grouped by manufacturers[19]. For each manufacturer, there may be multiple entries for the different instrument models available — keyboards, electric guitars, amplifiers, etc. For each model, there may be pictures, reviews, user manuals, and audio samples of how the instrument sounds.

Unlike most web traces, this web site was specifically configured to prevent caching, so the log represents all requests (not just the browser cache misses).

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

The data is based on a 1-week log file during 12-18 February of 1997 and the filtered data contains 8191 sessions and 892 web pages.

We divide the resulting set of transactions into a training (approx. 90%) and a testing set (approx. 10%) for experiments. Evaluation results showed that the hybrid recommender achieved more accurate predictions and relevant recommendations than using content-based or collaborative alone.

B. Evaluation Metrics

In order to evaluate the recommendation effectiveness for our method, we measured the performance of proposed method using two different standard measures, namely, Accuracy, Coverage [20]. Recommendation accuracy measures the ratio of correct, where correct recommendations are the ones that appear in the remaining of the user session. Recommendation coverage on the other hand shows the ratio of the pages in the user session that the system is able to predict before the user visits them; these metrics can be useful indicators of the system performance only when used in accordance to each other and lose their credibility when used individually [20].

C. Experimental Results

We evaluate our method under different settings. The first experiments were performed to evaluate system sensitivity to the size of visit window and recommendation window. We show the effect of them on efficiency of the proposed system. To consider the impact of visit window's size, we vary visit window's size from 1 to 12 and recommendation window's size from 1 to 20 (similar to previous system's experiment [8]).

We carried out experiments to investigate the performance of the proposed recommendation systems when providing different numbers of recommendations at different size of the visit window. The results are shown in figs. 1-3. Figs. 1-3 demonstrate the accuracy and coverage of our proposed approaches, content-based approach, collaborative approach and hybrid approach, respectively. In order to show the impact of size of the visit window (|w|) and size of the recommendation window (|w[`]| in these approaches, we changed these parameters, as shown in figs. 1-3.

Our evaluations indicate the best accuracy is achieved when using window sizes of 1 and the best coverage is achieved by using window size of 20.

We did a second experiment to further probe the relationship between size of recommendation window and system performance. Figs. 4-6 demonstrate the result of this experiment for various size of recommendation window. The results demonstrate that accuracy decreases when the size of recommendation window increases. The best performances are achieved when using size of recommendation window 1, and when we used recommendation window size larger than 5 results in lower accuracy. This is due to the fact that larger values of *w*' make weaker recommendations for the target user.

Also, in this figs, with the maximal number of recommendations increasing from 1 to 20, the coverage increases. When only one web page is recommended, the coverage is at its lowest, with a value of about 20%. If a maximum of 20 pages is recommended, the coverage fluctuates between 55% and 70%, depending on the size of the user window at which a particular recommendation is made.

Comparison the proposed systems indicated that the hybrid approach gains much better results than content and collaborative based approaches.

In the last part of our experiment, we consider performance of the proposed systems in comparison with a graph based recommender system discussed in [8]. Fig.7 has shown the comparison of the proposed systems performance with this system. Experimental results show that the our hybrid approach improves performance significantly and gains much better results than proposed content-based, proposed collaborative and navigation graph [8].

VII. Conclusions and Future Work

In this paper we have described a two-layer graph model that combines content-based and collaborative approaches. Using our model a recommendation becomes a graph search activity, and different graph search approaches can be applied. Our evaluation results show the flexibility of the proposed model to incorporate different sources of information to improve the quality of recommendations. Also results show that proposed model could significantly improve the recommendation effectiveness.

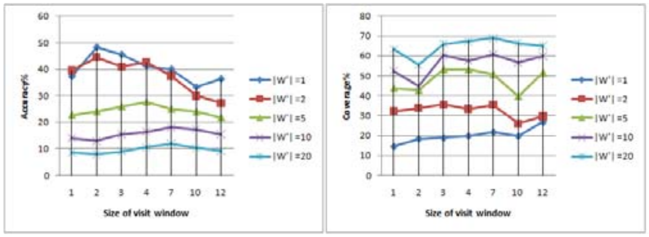

Figure 1. Average accuracy and coverage in content-based approach for various size of visit window

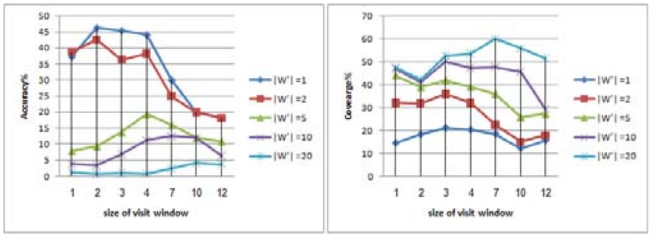

Figure 2. Average accuracy and coverage in collaborative approach for various size of visit window

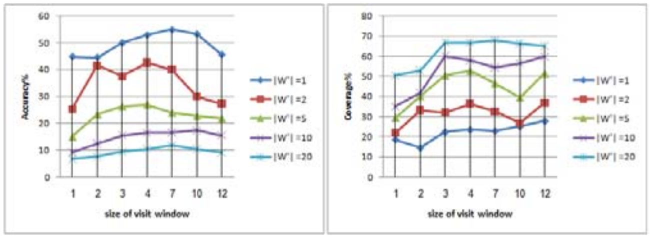

Figure 3. Average accuracy and coverage in hybrid approach for various size of visit window

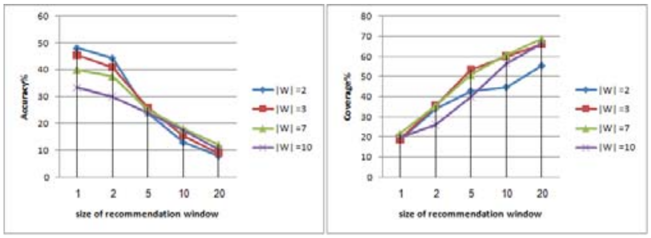

Figure 4. Average accuracy and coverage in content-based approach for various size of recommendation window

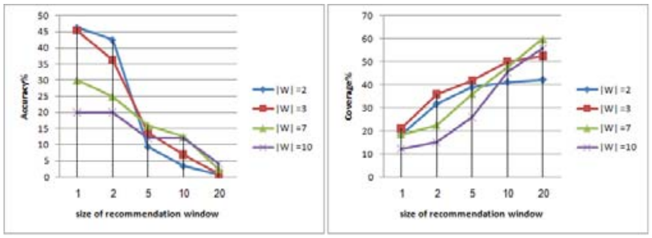

Figure 5. Average accuracy and coverage in collaborative approach for various size of recommendation window

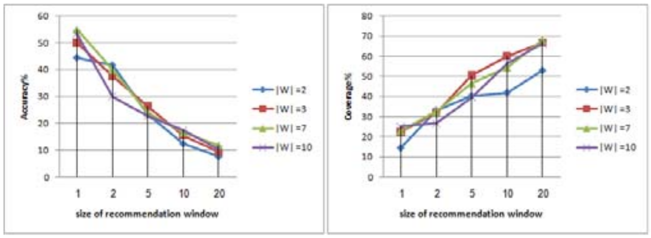

Figure 6. Average accuracy and coverage in hybrid approach for various size of recommendation window

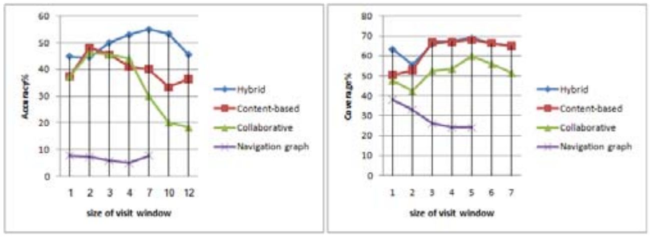

Figure 7. Comparing accuracy and coverage in content-based, collaborative, hybrid, and navigation graph[8]

REFERENCES

[1] R. Bruke, "Hybrid Recommender Systems," *School of Computer Science*, *Telecommunications and Information Systems*, *Springer Berlin Heidelberg*, 2007, pp. 377-408.

[2] B. Sarwar, G. Karypis, J. Konstan, J. Riedl, "Item-based Collaborative Filtering Recommendation Algorithms," *Proceedings of the 10th international conference on World Wide Web*, ACM New York, 2010.

[3] D. Lemire, "Scale and Translation Invariant Collaborative Filtering Systems," *Journal of Information Retrieval*, pp. 129-150,2003.

[4] C. Basu, H. Hirsh, W. Cohen, "Recommendation as Classification: Using Social and Content-based Information in Recommendation," *Proceedings of the 15th National Conference on Artificial Intelligence*,Madison, 1998.

[5] M. Bogdanov, M. Haro, F. Fuhrmann, E. Gomez, P. Herrera, "Content-based Music Recommendation Based on Use Preference Examples," *The 4th ACM Conference on Recommender Systems. Workshop on Music Recommendation and Discovery*, 2010.

[6] M. Goksedef, S. Gunduz-Oguducu, "Combination of Web Page Recommender Systems," *Journal of Expert Systems with Applications*, USA, 2010.

[7] C.C. Aggarwal, J.L. Wolf, K.-L. Wu, and P.S. Yu, "Horting Hatches an Egg: A New Graph-theoretic Approach to Collaborative Filtering," *In Proceedings of the Fifth ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, San Diego,1999, pp.201–212.

[8] Y. Wang, W. Dai, and Y. Yuan, "Website Browsing Aid: A Navigation Graph-based Recommendation System," *Journal Decision Support systems*, Elsevier Science Publishers B. V. Amsterdam, Netherlands, 2008.

[9] T. Liu, Y. Tian, and W. Gao, "A Two-phase Spectral Bigraph Co-clustering Approach," *KDD Cup and Workshop 2007, at the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2007.*

[10] I.S. Dhillon, "Co-clustering Documents and Words Using Bipartite Spectral Graph Partitioning," *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, California, 2001, pp.269–274.

[11] 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, pp. 3-23.

[12] Z. Huang, W. Chung, T.H Ong, and H. Chen, "A Graph-based Recommender System for Digital Library," *ACM/IEEE Joint Conference on Digital Libraries*, 2002, pp. 65-73.

[13] 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, pp. 259-274.

[14] R. Forsati, M.R. Meybodi, M. Mahdavi, A.G. Neiat, "Hybridization of K-means and Harmony Search Methods for Web Page Clustering," *IEEE International Conference on Web Intelligence and Intelligent Agent Technology*, 2008.

[15] W. Wang, O. R. Zaiane, "Clustering Web Sessions by Sequence Alignment," *13th International Workshop on Database and Expert Systems Applications*, University of Alberta, 2002.

[16] 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, pp. 1316-1330.

[17] V. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," Soviet Physics Doklady 10, 1966, pp. 707–710.

[18] B. Ziolko, J. Galka, D. Skurzok, and T. Jadczyk, "Modified Weighted Levenshtein Distance in Automatic Speech Recognition," *Krajowa Konferencja*, Department of Electronics, AGH University of Sience and Technology, 2010.

[19] M. Perkowitz, O. Etzioni, "Adaptive Web Sites: Automatically Synthesizing Web Pages," *In Proceedings of The Fifteenth National Conference On Artificial Intelligence*, 1998.

[20] M. Talabeigi, R. Forsati, M. R. Meybodi, "A Hybrid Web Recommender System Based on Cellular Learning Automata," grc, *2010 IEEE International Conference on Granular Computing*, 2010, pp. 453-458.