Journal of Academic and Applied Studies
(Special Issue on Applied Sciences)
Vol. 3(9) September 2013, pp. 1-12
Available online @ www.academians.org

Journal of Academic and Applied Studies

ISSN1925-931X

Create a Profile for User Using Web Usage Mining

Zeinab Khademali, Nima Attarzadeh, Mohammad Mehdi LotfiNejad
Department of Computer Science, Hendijan Branch, Islamic Azad University, Hendijan, Iran.

Abstract

In this paper we try to classify navigation patterns of web users automatically; therefore, a new method is presented in order to classify the user's navigation patterns and predict the user's future requirements. This method is based on the mining of web server logs. Thus, in order to build user's profile, a new method is introduced that, by registering user's setting and similarity measure of active user to neighboring users, constructs indices implicitly and brings them up to date based on created changes. In effect, we improve the performance of recommender engine by using navigation patterns of user and clustering similar users. Furthermore, we test the precision for different inputs in the model simulated based on the neural network and we determine that if, in registering user's profile, in addition to behavior history, current session is focused on as well, recommender engine will offer better results. This method that is based on user's navigation patterns is capable of offering the results from recommender engines based on user's requirement and interest. Advantage is evaluated based on two responsibilities: classification and prediction. The system has reached classification precision close to and prediction precision of about .This method can help web personalization and website better organization.

Keywords: **recommender engine, web usage mining, user's navigation patterns, neural network.**

I. Introduction

The big explosion of World Wide Web during the last 15 years has provided users with a lot of and increasingly developing information. Unfortunately, information explosion does necessarily not cause progress in the quality of our lives. Finding relevant knowledge and information among from a surfeit of available information can be very time consuming and even disappointing. Therefore, having an intelligent system capable of learning user's interests and filtering irrelevant interests automatically based on those interests or recommending corresponding information to user in a short time is important. Recommender systems deal with the problem of information overhead and help solving this problem by recommending items to users. Usually, there are a set of users and a set of items in a recommender system. Every user, U, gives some values to a set of items. This relation can be displayed by a matrix called user-item matrix. The responsibility of recommender systems is predicting the values that user U gives to a not-given-value item I or recommending some items to user U based on existing value-

giving's. In recommender system, techniques can be divided into three groups: collaborative filtering, content-based techniques and mixed method. Recommendations made by using collaborative filtering technique relies on giving value to item I by a set of users whose value-giving profile is the most similar to user U. content-based techniques consider some characteristics of item such as type, name, etc based on which make recommendations. Mixed methods make recommendations based on combining both techniques (Tuzhilin et al, 2005).

Currently, techniques of web usage mining are applied widely to discover the interests and navigation patterns of user from web server logs. Pattern mining (Zhou et al, 2004), association rule mining (Lin et al, 2000, Mobasher et al, 2001), and clustering (Mobasher, 1999, Phatak et al, 2002) discover different access patterns from web logs.

In order to analyze web pages, components such as modes and behaviors of user that are received by implicit or explicit factors should be taken into account. Out of the implicit factors are sequential and continuity of the user in searching pages, time passed for reading pages and date of reading pages, which denote searching behavior of the user and are used in classification. Of explicit factors special opinion poll forms and settings used by the user in registering his/her reactions can be mentioned; they are used in expressing the user's behavior and interaction (Chen et al, 2008).

Thus in this article we have dealt with introducing a new method that, by registering the user's settings and similarity measure of active user to neighboring users, constructs indices implicitly and brings them up to date based on created changes. Furthermore, we test the precision for different inputs in the model simulated based on the neural network and we determine that if, in registering user's profile, in addition to behavior history, current session is focused on as well, better result will be achieved through recommendation. Above-mentioned method uses information in log to determine user's requirements dynamically. Also we will evaluate profitability and advantage of the experimental system by using two defined standards: classification precision and prediction precision.

In this article, section 2 defines the basic concepts, in section 3 approaches and methods considered in several past decades are introduced, section 4 deals with the approach suggested by this article that builds implicitly user's profile and brings them up to date based on created changes by registering the user's settings and similarity measure of active user to neighboring user, and in section 5 the results obtained from testing the model simulated by using the neural network are evaluated.

II. main concepts

In this section a background that is important for understanding the method offered by this article is provided. First, web mining concepts are defined. Then personalization based on web usage mining is described. And, finally, clustering and neural network approaches are depicted.

i. web mining

Web mining is using data mining technique so that we could automatically be able to extract information from web services and documents. Web mining responsibilities can divide into four groups as follows (Kosala et al, 2004):

- Finding information: in this unit requested information is extracted and retrieved offline or online from textual documents existing in web.
- Selecting information and preprocessing: in this unit preliminary processing is performed automatically on the retrieved information.
- Generalizing: in this unit general patterns of personal websites are discovered.
- Analyzing: finally, this unit considers extracted patterns and estimates their validity. The patterns whose accuracy and validity are confirmed are accepted and offered.

Web mining can be divided into three main groups:

Web structure mining: it is a process that analyzes nodes and structural connections in a website by using graphs patterns.

Web content mining: it is a process that discovers effective and usable information from texts, photos, and audio and video data in web. Since textual data cover the largest part of data in web, web content mining is sometimes called web texts mining as well.

Web usage mining: it focuses on techniques that can predict the user's behaviors while interacting in web. The main responsibility in web usage mining is retrieving useful and meaningful information from profile of using web servers according to user's searching, which, in turn, is divided into three steps: preprocessing data, discovering patterns and analyzing patterns.

Preprocessing step includes processing the behavior of site's files and user's profile data in pages, sites and files of server session. Patterns discovering step includes processing server session files to rules, patterns and statistical information. Patterns analyzing step deals with rules and statistical information obtained from patterns discovering step.

Preprocessing step, in turn, is divided into three subgroups of preprocessing of content, preprocessing of structure and preprocessing of usage. Preprocessing of content classifies site's files in the form of different pages to help the analysis of patterns. Preprocessing of structure is a step that turns site's files into topology and site's correlation in order to help identifying pages. Finally, preprocessing of usage turns used raw data into a flow of user's behavior in the form of data cleaning, user's identification, and session's identification processes.

ii. web usage mining for personalization

Web usage mining is one of the applications of data mining technique in order to use log files for improving web site designing (Cooley et al, 1999). Log files of web servers potentially include useful empirical data to improve websites performance and content some advantages for some applications, especially in business cases. By analyzing these files predicting those links that have a positive effect on increasing the performance of website and are very useful for websites designers will be possible (Yang, 2005). For example, predicting links can play a very effective role in loading documents that may be visited by a visitor while he/she is reading a current page. By using log file of work flow, some strategies can be offered to solve the problems that hinder

improving business processes (Subramaniam, 2006). It actually focuses on techniques that can predict user's behaviors while interacting in web. The main responsibility in web usage mining is retrieving useful and meaningful information from log file of usage web servers based on user's behaviors, which, in turn, is divided into three steps of preprocessing data, discovering patterns and analyzing patterns.

iii. clustering

The set of input patterns $X=\{x_1,x_2,\ldots,x_n\}$ includes n objects. Each of these objects is tantamount to a vector with a length of s of characteristics. These objects should be clustered in K groups, namely $C= \{c_1,c_2,\ldots,c_k\}$, that do not overlap each other (Tuzhilin et al, 2005).

In this article, k-means algorithm is used for clustering similar users. In spite of its simplicity, this method is considered to be a basic one for many other clustering methods such as fuzzy clustering.

iv. neural network

Artificial neural networks are a kind of simplified system of real nervous systems that are applied frequently in solving different problems in science. The most important advantage of these networks, in addition to the simplicity of their usage, is their high ability.

Artificial neural network is an idea for processing the information inspired by biological neural system, and it processes the information just like the brain. This system is composed of a lot of processing elements called neurons that act together compatibly to solve a problem.

III. related works

In order to control the information existing in web, researchers have tried to retrieve the most relevant existing information so that they could increase the content and richness of the retrieved information by using approaches that are inclined to analyzing the relation between the question asked and the answer retrieved, thereby achieve users absolute or relative satisfaction.

(Mohammadi dostdar et al, 2011) a combined recommender system has been introduced that combines the results obtained from two approaches of content and collaborative filtering based recommender systems in terms of a two-layer graph design and conducts this combination based on graph partitioning. This graph consists of users' layer and web pages layer and embraces user-user, web page-web page and user-web page relations. Each node in web pages layer denotes a web pages, each node in users' layer denotes a user.

Hereby the similarity among web pages and the similarity among users are obtained and graph partitioning is used for classifying users and web pages. The measures proposed in this article are coverage and precision. The precision of recommendations is equal to the ratio of true recommendations to total recommendations. True recommendation means the recommendation that has been made based on the visited part of a user's session and will happen in the continuation of his/her session. The coverage of recommendation measures the ability of system in predicting all pages that may be intended by users.

(Forsati et al, 2009) a mixed algorithm of the information of users' searching and the link among web pages to recommend pages to users have been developed. The measure introduced for calculation the weight of pages visited by users' uses duration of visiting a page and frequency of visiting it. In the presented algorithm, the first page is suggested by using proposed weighted association rules. Then that page is extended by HITS algorithm and the pages that along with that page belong to the same classification. For the classification of pages an algorithm has been introduced based on the learning automata and graph partitioning algorithms. Parameters that affect the performance of algorithm are the size of recommendation window and recommendation thereshold.

(Nicholas et al, 2006) studied the behavior of information seeking of the users by using web mining techniques. They found out that most web users hadn't read web pages for long and before leaving web resources they had just browsed through a limited number of web pages and items. The two meters considered in this article refers to the number of items visited in the current session and the number of visits as well. [Breeding, 2005] studied the behavior of information searching of the users more profoundly by using a certain group of websites' users and web logs' analysis software. He didn't limit himself to investigating the behavior of information searching of individual users; rather he studied a group of them through sessions held by them.

[Nicholas et al, 2005] by studying transactions of searching weblogs, maintained that measuring tools that were discovered from those resources were useful means for studying the rate of performance and also the degree of satisfaction or dissatisfaction of search engines. They took advantage of two factors of measuring time passed between searching sessions and the number of searches performed in each session in order to study the information searching behaviors of the users of search engines. Another study, at the same level, was conducted by [Nicholas et al, 2006] that investigated the information searching behavior of the users of digital scholarly journals. Their concentration was on the users of Blackwell Synergy database, and they used the factor of the number of held sessions and visited and requested items to investigate the information searching behavior of the members of scientific board of above-mentioned base's digital journal. These researchers presented that if that kind of research had been completed with qualitative investigation of the information searching behavior of the users, better results would have been obtained.

The resource of data in those studies was all the pages visited by site's customers in a log file on web server. The analysis of these data files helps the evaluators of sites to identify the main point of the course of heavy-traffic levels in site. Searchers can extract the identify of site and pages visitors and a section in a site that has been visited by visitors.

IV. recommended method

Since profiles in server are stored orderly and are not specific to one user, but they can used by different users, and also for every user the information of different searches is maintained, the retrieving of the information is often accompanied with error. Therefore, these data should be preprocessed and prepared before one is able to use them.

Thus, in this system, for extracting user's sessions from web server log, we start first by the preliminary preprocessing of web server logs. The preprocessing of logs includes data cleaning, user identification and user's session's identification.

1. **data cleaning**

In web server records just those records that have relevant information are saved and maintained. In order to eliminate inappropriate records from log file, data cleaning is used which includes the elimination of the following data:

- Requests carried out by automatic programs
- Requests for audio files
- Records with unsuccessful HTTP statues codes
- Log's records through requested methods except Post or Get

After data cleaning, users are identified based on IP and ID addresses and after that user's sessions are identified.

2. **session identified**

A session of activities performed by user covers a moment when he/she has entered the site till the moment when he/she has left it. A threshold is considered for the duration of session, and if it passes a certain limit, it denotes that there is another access session of user. Experience shows that a 30-minutes threshold has been suggested for the duration of session [Spiliopoulou et al, 2003].

Sessions identified by two sets that are dependent on the date of log's records are divided into two sets: educational and experimental. By using educational set, user's profiles can be created in usage mining, whereas experimental set is prepared for classification and prediction experiments.

3. **constructing vector of session**

After preprocessing step, sessions should be classified into clusters, depending on their common characteristics; navigation pattern index of users can be obtained through their clustering. Then these patterns help user with making profile.

User's session can be depicted as a vector of the weight of the pages that has been visited during a certain time period. Therefore, user's session can be described as follows:

In order to evaluate the degree of user's interest, his/her sessions should be extracted first, and, then users should be clustered based on the degree of their similarity. When a new user is added, the degree of his/her similarity to clusters is investigated and his/her interests are determined, depending on which cluster he/she belongs to. Suppose P is the total set of pages accessible by users, as follows:

$P=\{p_1,p_2,\ldots,p_m\}$

Where each p_i is calculated by a unique url.

Also, set S shows user's access sessions as a subset:

$S=\{s_1,s_2,\ldots,s_n\}$

Where each s_i is a subset of P. each session, in effect, is a sequence of pages that have been visited in certain duration; in article, this duration is usually considered 30 minutes. To construct vector of users' sessions web usage mining (WUM) is used. Each session is an m-dimension vector as follows:

$S_i=\{W(P_1,S_i),W(P_2,S_i),\ldots,W(P_m,S_i)\}$

Where the weight of each page, P_j, is determined in ith session. The weight of each page signifies the degree of the interest of user. In order to calculate the weight, and actually the degree of the user's interest in page, two factors of "frequency of page" and "duration of page view" are used.

In the following equations it is indicated that how frequency and duration can be computed.

$$frequency(page) = \frac{Number\ of\ visits(page)}{\sum_{page \in visited\ pages}(Number\ of\ visits(page))} \quad (1)$$

$$\text{Duration(page)} = \frac{Total\ Duration(page)/Lenght(page)}{Max_{page \in visited\ pages}\ (Total\ Duration(page)/Lenght(page))} \quad (2)$$

The importance of total page can be getting by combining the two above-mentioned criteria. In this system, harmonic mean of frequency and duration has been applied to indicate the degree of the user's interest in a web page in session:

$$\text{Interest(page)} = \frac{2 * Ferequency(page) * Duration(page)}{Ferequency(page) + Duration(page)} \quad (3)$$

Finally, for each session there will be a vector as follows:

$S_i= \{W_1, W2,\ldots, Wm\}$

Where W_i denotes the weight of ith page in a certain session. The number of m dimensions should not excess a reasonable amount; therefore, pages whose support is low or high should be cleaning.

4. Constructing user profile

Every user has K session in each of which he/she has visited a sequence of pages such that S_1, S_2, …, S_k is a set of sessions of ith user. Mean vector S_{ui} is considered user's index or interest, u_i. The weight of page on mean vector is obtained from the mean weight of that page in all user's sessions. When calculating mean vector of session, the history of the user's behavior is

considered too. In order to achieve a better result, the user's partial session can be used, in addition to his/her behavior history.

- **the role of behavior history in setting profile**

Every user's profile is usually derived from different sources and content information about various aspects of the behavior of user. After the preliminary profile of user is formed, the user will make his/her request, and the system will retrieve the results in accordance with made request and similarity and closeness of the request to the neighboring profile. According to the retrieved results, user can select a web page from among retrieved pages or search more web pages through existing links in order to meet his/her information requirements. The system monitors behavior history of the user and modifies the user's profile with every change. Next times, when the user makes a request, results will be constructed based on the new profile.

5. clustering profiles

Now, mean vectors of sessions should be compared and clustered based on their similarities. So that there will be a subset of similar user's profiles in each cluster. Clustering is a crucial step in our method because clusters determine neighbors of the destination user. Clustering algorithms are divided into two main groups: classical method and heuristic method. K-means algorithm has been used for clustering. This is one of the most important and popular classical algorithms whose performance is strongly dependent on original state of question and initial centers of clusters. In this algorithm the number of clusters (K) should be inserted into algorithm as an input parameter and manhatan distance should be used for calculating the distance between two objects in clustering. Set of clusters is as follows:

$C=\{c_1,c_2,\ldots,c_k\}$

K is the number of clusters. As the representative of each cluster, the mean of each cluster, m_c, is computed that displays navigation pattern of the users of a cluster in a certain set of accessed web pages. Finally, as the result of clustering profiles, a set will be given as follows:

$NP=\{np_1,np_2,\ldots,np_k\}$

Where each p_i is a subset of the set of web pages P.

6. constructing recommender system by using neural network

In this step it has been tried to create, by receiving the user's current session, a recommendation list for him/her. From neural network, the most similar cluster to the user's session is found and suitable pages are suggested to him/her. By using navigation patterns obtained in the preceding step the network is instructed in which each navigation pattern is regarded as an input of neural network and output of the network is the number of the cluster that has already been determined for each navigation pattern.

After instructing neural network, as a new user enters the site, first the user's current session should be prepared in a manner so that it would be suitable as the input of neural network. As a matter of fact, for user's session, a profile is created based on the weight of pages. Now it should be determined that current session profile belongs to which navigation pattern, i.e. to which

cluster. To do so, current session profile is given to the input of neural network and the network will determine suitable cluster for session. After the number of cluster is determined, those pages of cluster that have not been visited in current session have a high potential of being the next pages that user is interested to visit; therefore, they will be interested in the recommended list.

V. conclusion

In this paper, in order to provide useful information required by user, we suggested an approach for recommending web pages to the current user based on his/her profile. In effect, we suggested an approach based on navigation patterns of user that are results returned by engine recommending in accordance with user's requirement. Recommended approach make a better search possible compared with earlier approaches by registering changes in settings of each user. This approach separate pages relevant to user's interests from those pages that are irrelevant. In order to investigate the effect of the new approach, we conducted research on constructing user's profile based on history of their behavior. Also, in another research we aimed at setting user's history implicitly and set user's profile without their intervention. In this way, we could have brought profile up to date, if interests and settings of each of users had changed. Regarding the fact that each user's profile is not constructed only based on his/her own settings, we studied search activity of the user in a day in more detail. In adaptive search for each user if the focus is on user's current session rather than user's search history, better results will be gotten. Moreover, a clustering technique was recommended according to k-means technique for adapting the results of recommender engine to user's requirement. This method used similarities among users to identify neighbors of destination user. By using k-means clustering similar users groups were identified.

Based on the results obtained from this study it can be concluded that recommended approach, through considering the history of user's behavior in constructing profile, can modify and bring up to date user's profile in case his/her settings change. In the history of behavior if the focus is more on the user's current session, recommender engine will have a higher performance, and user will receive more relevant pages, in a shorter time, for his/her behavior. In other words, we offered a system of recommending usage-based web that dynamically suggested pages in which user was interested by using computational intelligent techniques. This system uses neural technique to determine the classifications of similar and common interests of users. In comparison to systems based on associative rules, the below results are obtained for precision and coverage:

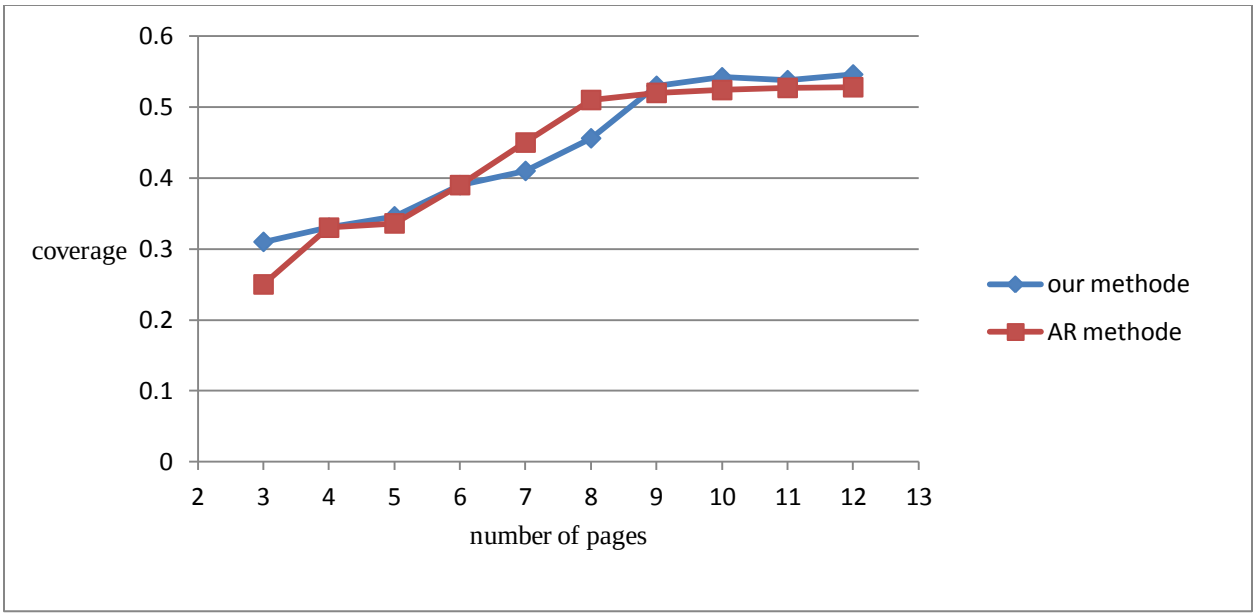

Figure1: coverage of the recommendations

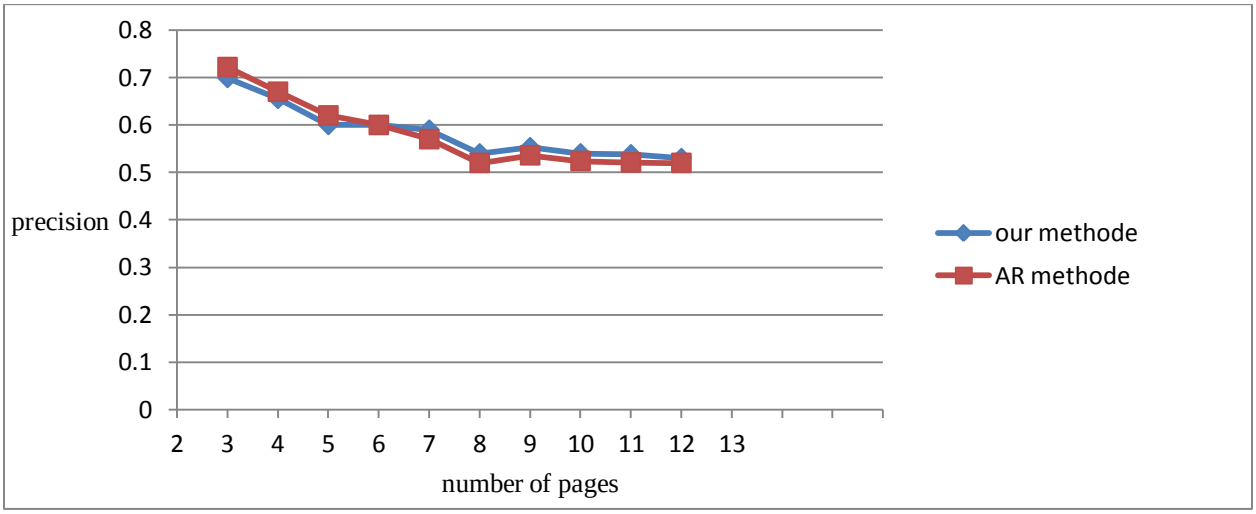

Figure2: precision of the recommendations

- **future work**

As it was mentioned, recommended method of this study emphasized the construction of user's profile; however, because textual data are more than other kinds of data in web, it just considers textual data in user's profile. If networks with a high broadband are generalized widely, spreading information in its different forms such as music, audio, video, etc. won't be beyond our expectation. We will also try to consider a similarity standard in resultant clusters so that we could be able to calculate recommended quality that has been provided by other users. Furthermore, we are going to attribute users to several clusters (overlapped clusters) and use

these clusters for recommendation since in real situations also people usually have different interests.

References

Berendt B, Mobasher B, Spiliopoulou M and Wiltshire J (2001), Measuring the Accuracy of Sessionizers for Web Usage Analysis, Proceedings of the Web Mining Workshop at the 1st SIAM International Conference on Data Mining, PP 7-14.

Breeding M (2005), Analyzing Web Server Logs to Improve a Site's Usage, Computers in Libraries, Vol(25), Number 9, PP 26-29.

Chen P.Z, Sun C.H and Yang S.Y (2008), Modeling and Analysis the Web Structure Using Stochastic Timed Petri Net, journal of software, Vol(3), Number 8, PP 19-26.

Cooley R, Mobasher B and Srivastava J (1999), Data Preparation for Mining World Wide Web Browsing Patterns, Knowledge and Information Systems, Vol(1), PP 5-32.

Forsati R and Meybodi M.R (2009), Algorithmic Based on Structure Pages and User's Usage Information for Recommending Web Pages, the 2th Iran data mining conference.

Kosala R and Blockeel H (2000), Web Mining Research: a Survey, ACM SIGKDD Explorations Newsletter, Vol(2), Issue 1, PP 1-15.

Lin W, Alvarez S.A and Ruiz C (2000), Collaborative Recommendation via Adaptive Association Rule Mining, Data Mining and Knowledge Discovery, Proceedings of the Web Mining for E-Commerce Workshop (WebKDD'2000).

Mobasher B (1999), A Web Personalization Engine Based on User Transaction Clustering, Proceeding of the 9th Workshop on Information Technologies and Systems (WITS'99).

Mobasher B, Dai H, Kuo T and Nakagawa M (2001), Effective Personalization Based on Association Rule Discover from Web Usage Data, WIDM '01 Proceedings of the 3rd international workshop on Web information and data management, USA, PP 9-15.

Mohammadi dostdar H, Forsati R and Meybodi M.R (2011), Recommender System of Combined Web Based on 2-layers Graph and Partition of Graph, the 5th Iran data mining conference.

Nicholas D, Huntington P, Jamali H.R and Watkinson A (2006), the Information Seeking Behaviour of the Users of Digital Scholarly Journals, Information Processing & Management, Vol(42), Issue 5, PP 1345-1365.

Nicholas D, Huntington P and Watkinson A (2005), Scholarly Journal Usage: the Results of Deep Log Analysis, Journal of Documentation, Vol(61), issue 2, PP 248-280.

Nicholas D, Huntington P, Jamali H.R and Tenopir C (2006), Finding Information in (Very Large) Digital Libraries: A Deep Log Approach to Determining Differences in Use According to Method of Access, journal of Academic Librarianship, Vol(32), Number 2, PP 119-126.

Phatak D.S and Mulvaney R (2002), Clustering for Personalized Mobile Web-Usage, Proceedings of the IEEE International Conference on Fuzzy Systems, PP 705-710.

Spiliopoulou M, Mobasher B, Berendt B and Nakagawa M (2003), A Framework for the Evaluation of Session Reconstruction Heuristics in Web-Usage Analysis, Informs Journal on Computing, Vol(15), Issue 2, PP 171-190.

Subramaniam S (2006), Optimizing Business Processes Through Log Analysis, University of California, Riverside.

Tuzhilin A and Adomavicius G (2005), Towards the Next Generation of Recommender Systems:A Survey of the State-of-the-Art and Possible Extensions, IEEE Transactions on knowledge and data engineering, Vol(11), Number 6, PP 134-147.
Yang Z (2005), Web Log Analysis: Experimental Studies.
Zhou B, Hui S.C and Chang K (2004), An Intelligent Recommender System Using Sequential Web Access Patterns, IEEE Conference on Cybernetics and Intelligent Systems, Vol(1), PP 393-398.