http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 9 (4): 790-796, 2010
ISSN 1812-5638

Mining Web Navigation Profiles For Recommendation System

Y.M. AlMurtadha, Md. N.B. Sulaiman, N. Mustapha and N.I. Udzir
Department of Computer Science, Faculty of Computer Science and Information Technology,
University Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

Abstract: This study explores web usage mining, for which many data mining techniques such as clustering, classification and pattern discovery have been applied to web server logs. The output is a set of discovered patterns which form the main input to the recommendation systems which in return predict the next web navigations. Most of the recommendation systems are user-centered which make a prediction list to the users based on their long term navigation history, user's databases or full user's profiles. Companies wish to attract anonymous users, directed them at the early stages of their visits and get them involved with their websites. Learning and mining the web navigation profiles followed by enhanced classification to the similar activities of previous users will provide an appropriate model to recommend to the current anonymous active user with short term navigation. Using CTI dataset, the experimental results show better prediction accuracy than the previous works. An adaptive profiling to save time is a key factor for future works.

Key words: Usage profiling, web usage mining, recommender systems

INTRODUCTION

The rapid growth of the Web in the last decade makes it the largest publicly accessible data source in the world. The web now becomes one of the main sources of the information. Now it includes more than 4 billion pages, with about 1 million added every day (Markov and Larose, 2007). Getting information from the internet is like drinking water from the fire hose. Web mining aims to discover useful information or knowledge from Web hyperlinks, page contents and usage logs (Liu, 2007). Yet, an important problem is how to mine complex data formats including Image, Multimedia and Web data (Yang and Wu, 2006). Based on the primary kinds of data used in the mining process, web mining tasks can be categorized into three main types: Web structure mining, Web content mining and Web usage mining (Liu, 2007). Web structure mining discovers knowledge from hyperlinks, which represent the structure of the web. Web content mining extracts useful information/knowledge from Web page contents. Web Usage Mining (WUM) mines user access patterns from usage logs, which record clicks made by every user. The goal is to capture, model and analyze the behavioral patterns and profiles of users interacting with a Web site. The discovered patterns are usually represented as collections of pages, objects, or resources that are frequently accessed by groups of users with common needs or interests. These patterns are discovered by applying some clustering algorithms on the preprocessor phase of the web usage mining and classification algorithms on the web mining process. The output of the WUM is some patterns that may be the input to the Recommendation systems Engine which is one of the application areas of the Web usage and gives the ability to predict the next visited page for a given user. Recommender system alleviates the information overload by pruning the information spaces and directing users toward the items that best represents their interests (Taghipour *et al*., 2007).

The major application areas for WUM fall into 5 categories: personalization, system improvement, site modification, business intelligence and usage characterization (Srivastava *et al*., 2000). Recently, many researches tried to improve the prediction accuracy of the recommendation systems; however, the current recommendation systems can not satisfy the users especially with the increasing growth of the web sites, the increasing number of the web users and the preferences changing of these users at any time (Jalali *et al*., 2008) and most of them are user-centered recommendation systems. Recommender systems have been developed using various approaches and can be categorized in various ways (Burke, 2002). Collaborative filtering is the most widely used technique in these systems (Deshpande and Karypis, 2004). Reinforcement Learning has been used for recommendation in several applications (Joachims *et al*.,

Corresponding Author: Yahya M.A. Murtadha, Department of Computer Science,
Faculty of Computer Science and Information Technology, University Putra Malaysia,
43400 UPM Serdang, Selangor, Malaysia

1997; Taghipour *et al*., 2007). Web usage mining techniques have been widely used to build recommendation systems. Analog (Yan *et al*., 1996) is structured according to an off-line component to build session clusters and an online component to build active user sessions which are then, classified according to the generated model. The classification allows returning the requested page with a list of suggestions related to the ones in the active session. The geometrical approach used for clustering is affected by several limitations, related to scalability and to the effectiveness of the results found. Liu and Kešelj (2007) proposed a novel approach to classifying user navigation patterns and predicting user's future requests based on the combined mining of web server logs and the contents of the retrieved web pages. Baraglia and Silvestri (2004, 2007) proposed a WUM system called SUGGEST, that provide useful information to make easier the web user navigation and to optimize the web server performance. Potential limitation of this architecture might be: (a) the memory required to store web server pages is quadratic in the number of pages. This might be a severe limitation in large sites made up of millions of pages (b) it does not permit us to manage Web sites made up of pages dynamically generated. Mobasher *et al*. (2000a) presented WebPersonalizer, a system which provides dynamic recommendations as a list of hypertext links, to users (Mobasher *et al*., 2000a, b; Nakagawa and Mobasher, 2003) presented systems that take the advantage of combining content, usage and structure are introduced. Mobasher *et al*. (2002) presented and experimentally evaluate two techniques, based on clustering of user transactions and clustering of pageviews, in order to discover overlapping aggregate profiles that can be effectively used by recommender systems for real-time Web personalization. The prediction engine has to make a recommendation list to the user session from multiple profiles based on match score that must exceed the threshold. Jalali *et al*. (2008) proposed a novel approach based on LCS algorithm for classifying user navigation patterns for predicting users' future requests. All of these works attempt to find architecture and algorithm to improve accuracy of personalized recommendation, but the accuracy still does not meet satisfaction. Also, almost all of them are user-centered prediction engine which concentrate on recommending the next visited pages based on long term previous navigation of the user.

MATERIALS AND METHODS

An accurate prediction for the new visiting page will enable the companies to attract and direct the unidentified user based on their early stages navigations. The aim of the experiments carried out on 2009 was to evaluate the ability on predicting the new visiting page to the current active navigation session for any unidentified user using CTI dataset for the testing purpose. Here, we will describe the methodology of learning and mining the navigation profiles and the prediction evaluation.

Following the standard data mining process (Fayyad *et al*., 1996) the overall web usage mining process can be divided into three inter-dependent stages: data collection and pre-processing, pattern discovery and pattern analysis. In the pre-processing stage, the clickstream data is cleaned and partitioned into a set of user transactions representing the activities of each user during different visits to the site. In the pattern discovery stage, statistical, database and machine learning operations are performed to obtain hidden patterns reflecting the typical behavior of users, as well as summary statistics on web resources, sessions and users. In the final stage of the process, the discovered patterns and statistics are further processed, filtered, possibly resulting in aggregate user models that can be used as input to applications such as recommendation engines. As shown in Fig. 1, the proposed architecture consists of two main components, namely the offline and online. In the offline component two steps are taken. The first one is clustering the filtered sessionized pageviews into clusters of similar pageviews. The second step is learning the profiles based on the preformed clusters. The online component is responsible for matching the new user request (current active session) to the profile shares common interests to the user.

The offline component: Here, we describe the two steps required for the offline component, namely clustering and the learning the navigation profiles.

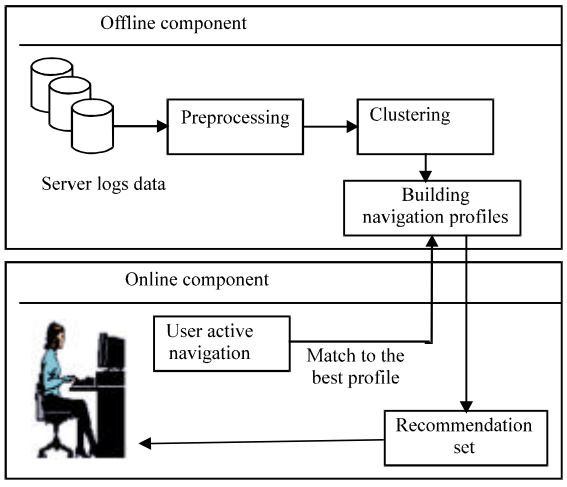

Fig. 1: The proposed recommender system overview

Table 1: Precision's sample for different k-clusters

k	Recommendation precision
10	0.627584
20	0.614564
30	0.59648
40	0.55

Sessions clustering: One important use of clustering in web usage mining is aimed at finding groups which share common interests and behaviors by analyzing the data collected in Web servers. We used K-Mean clustering algorithm to cluster the preprocessed and filters web server logs with different K-values. For the clustering purpose, we used CTI.std file as an input to the K-Means clustering algorithm. The file contains 13745 sessions with 682 pageviews visited by different users. The file represents a session-pageview matrix where each column is a pageview and each row is a session represented as a vector. The entries in the Table 1 correspond to the amount of time (in sec) spent on pageviews during a given session. The pageview durations were maxed out at 999 sec. For each session, the pageview duration of the last pageview in that session, was estimated to be the average duration of that pageview across all sessions (in which the pageview does not occur as the exit page). The output is K-clusters (i.e., 10 clusters) each contains sessions with similar pageviews. These clusters will be the input to the next step (building the navigation profiles). Table 1 shows different accuracy evaluation for different k values. This is true because the cluster's size does matter for calculating the confidence support and the mean weights.

Learning the navigation profiles: The discovery of patterns from usage data by clustering the web transaction into clusters of user sessions or pages, by itself is not sufficient for performing the personalization tasks (Mobasher *et al*., 2002). The critical step is the effective derivation of good quality and useful navigation profiles from these patterns. The discovery of aggregate usage profiles or patterns through clustering, as well as other web mining techniques, have been explored by several research groups. However, in all of these cases, the frameworks proposed have not been extended to show how these profiles can be used as a part of a recommender system (Mobasher *et al*., 2002). Normally, users cannot be identified from anonymous web server logs due to many reasons like using of a single computer by multiple users and dynamic IPs. According to (Suryavanshi *et al*., 2006) these anonymous users that every company wishes to attract and get involved with its website when they first visit the site which means only the current active navigation is needed rather than a full user profile. We used the clusters produced by the clustering step (previous step) to build the usage or the navigation profile with one profile for each cluster by setting the min_sup and min_weight. The navigation profile contains only those pageviews that passed certain confidence support and weights values. The confidence support determines the frequency occurrence on those pages in the cluster. The min_sup values are used to filter out profile elements which do not have sufficient support while min_weight values are used to filter out profile elements which have low average weight (navigation time spent visiting this page). We stress again that these profiles do not consider specific users since, we do not take the full users history (long term web navigation) in account during obtaining the profiles.

To summarize, we construct a navigation profile as a set of pageview-weight pairs:

$$\text{profile} = \{ p, \text{weight}(p) \mid p \in P, \text{weight}(p) \geq \text{min\_weight} \}$$

where, P ={p1, p2, . . . , pn}, a set of n pageviews appearing in the transaction file with each pageview uniquely represented by its associated URL and the weight (p) is the (mean) value of the attribute's weights in the cluster.

Figure 2 shows a navigation profile database snapshot of two profiles obtained for two clusters 1 and 2 where, each profile contains related pageviews. For example, profile 1 represents the activity of a user interested in the courses and the programs offered while, profile 2 represents the activity of a user interested in the pageviews related to the admission and advising.

The online component: After the navigation profiles are extracted from the previous sessions, many preprocessing steps are to be taken. First, the weights are normalized so that the maximum weight in each navigation profile is 1. Then, all the profile's pageviews are sorted in descending order according to their weights. Finally, all the highly frequent pageviews like the index pages are removed. The online component then is ready to assign the current navigation activity to the best profile among the navigation profiles extracted by the offline component. When, the user navigates the internet, the web server will start to keep his logs on a file. This file can be accessed to extract the current active navigation web pages called the active session. Using this active session, the online component is responsible for assigning this user activities to the best navigation profile where by a recommendation list is to be created from those pageviews not visited by the user and attached to the user navigation list. Two sequences methods are applied to enhance and assure the classification. First, statistical classification aimed to

ProfileNo	Weight	PView
1	0.4969	/cti/studentprofile/studentprofile.asp?section=mycti
1	0.4306	/courses/
1	0.323	/courses/syllabilist.asp
1	0.3184	/programs/
1	0.2504	/authenticate/login.asp?section=mycti&title=mycti&urlahead
1	0.1543	/people/
2	0.575	/admissions/
2	0.7	/cti/studentprofile/studentprofile.asp?section=mycti
2	0.6275	/cti/advising/display.asp?page=intranetnews
2	0.1644	/cti/advising/login.asp
2	0.1118	/people/
2	0.1009	/authenticate/login.asp?section=mycti&title=mycti&urlahead
2	0.08	/advising/

Fig. 2: Examples of Navigation profiles database

assign the active session to the best profile with the highest number of matched pageviews. Second, use the cosine coefficient to find the similarity with the profiles that may meet or be missed using the first method. Finally, make a recommendation list based on these selected profiles from those pageviews that pass a certain threshold. To do so, break each coming active session into its pageviews and then classified it to the best profile with the highest number of matched pages. If two or more profiles equal in the number of matched pages, then both are taken for the prediction. In the testing phase, the active session will be divided to two parts: surrogate session with sliding window size n (in the experiments we used n = 3) and the remaining for the comparison purpose. Only active sessions with equal to or larger than five pageviews are used which is the average size of the dataset. Figure 3 shows the minimum number of pages per active sessions used for the test purpose where, x axis represents the session and Y axis represents the number of pages.

Since, both the active session and the choose profile can be represented as vectors; the cosine coefficient commonly used in information retrieval was used to do the matching purpose.

$$\text{profileMatch} = \frac{\sum_i w_i^c \cdot p_i^p}{\sqrt{\sum_i \left(w_i^c\right)^2 \times \sum_i \left(p_i^p\right)^2}} \quad (1)$$

where, w_i^c is the associate weight for the corresponding pageview reference in the active session in binary values (0 for absence and 1 for presence) and p_i^p is the associate weight for the corresponding pageview reference in the profile.

A recommendation score is computed for those items not already visited by the user in the active session in order to recommend them based on their scores.

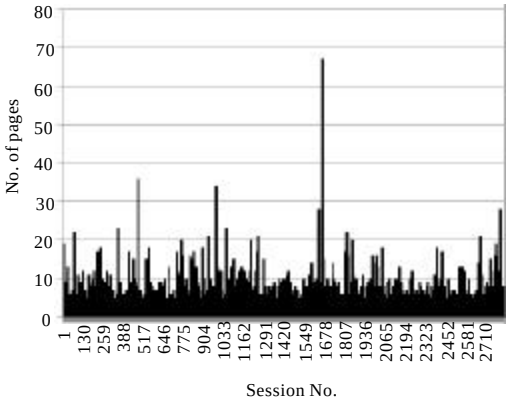

Fig. 3: Minimum number of pages in a session

$$\text{RecScore} = \sqrt{[\text{PageWeight} \times \text{ProfileMatch}]} \quad (2)$$

If the recommendation score is higher than the recommendation threshold, then select it. Various values from 0.1 to 1.0 are taken for the recommendation threshold. According to Mobasher *et al.* (2002), two factors are used in determining this recommendation score: the overall similarity of the active session to the profile as a whole and the average weight of each item in the profile computed during learning the profiles (offline component).

EXPERIMENTAL DESIGN

Our experiments have been conducted on DePaul University CTI logs file dataset which contains the preprocessed and filtered sessionized data for the main DePaul CTI Web server (http://www.cs.depaul.edu). The data is based on a random sample of users visiting this site for a 2 week period during April of 2002. The original

Fig. 4: The cti.tra filtered sessionzed data in transaction format

(unfiltered) data contained a total of 20950 sessions from 5446 users. The filtered data files were produced by filtering low support pageviews and eliminating sessions of size 1. The filtered data contains 13745 sessions and 683 pageviews. Based on the proposed architecture, a recommendation system is developed using Microsoft VC++ connected to Microsoft Access database through an Open Database Connection (ODBC). We used CTI dataset which contains 13745 sessions with 683 pageviews for the experiments with 75% for training and 25% for testing. It contains 4 files. Cti.cod, cti.tra, cti.std and cti.nav. As mentioned before, we used the cti.std file to generate the clusters and building the navigation profiles. The session forms the rows of this file and the pageviews form the columns. For the testing purpose, we used cti.tra which contains the filtered sessionized data in transaction format. As Fig. 4 shows, each line in this file corresponds to the sequence of pages visited during one session. There is a direct correspondence between the rows in this file and the rows in the file cti.std. While, the order of occurrence of pageview in each session represents the order in which these pageviews were visited, the transactions do not contain repeated visited to the same pageview in the same session. Thus, only the first access to a pageview is recorded as part of the transaction.

EXPERIMENTAL EVALUATION

We used the precision, coverage and F1 standard measures in order to evaluate the recommendation effectiveness. Assume that, we have active current session A taken from the evaluation set and we have R as a recommendation set using the prediction engine over the navigation profiles. W represents the items that already visited by the user in A. The precision is defined as:

$$\text{Precision}(R,A) = \frac{|R \cap (A - w)|}{|R|}$$

and the coverage is defined as:

$$\text{coverage}(R,A) = \frac{|R \cap (A - w)|}{(A - w)}$$

Finally, F1 is defined as:

$$F1(R,A) = \frac{2 \times \text{Precision}(R,A) \times \text{coverage}(R,A)}{\text{Precision}(R,A) + \text{coverage}(R,A)}$$

RESULTS

Due to many reasons, like using of a single computer by multiple users and dynamic IPs, users cannot be identified from anonymous web server logs. When they first visit the site, the web server records these anonymous users current active navigation rather than a full user profile. Every company wishes to attract these anonymous users and get them involved with its website, which means only the current active navigation is needed rather than a full user profile. Online shopping website's prediction engine must predict the user's new purchase (registered and non-registered users) and offer it to them based on their profile. The website's prediction engine does not has a full navigation history (cookies, navigation DB, etc.) composing the user's profile. The window sliding over the current active session with size n helps as a short term history for the anonymous user. For every transaction t (represents an anonymous user), the last n visited pages will be taken as an active session and the prediction engine will generate a recommendation set to this session. The recommended set will be compared to the remaining set of the transaction t. Table 2 shows the recommendation set contain pages recommended from the profile No. 5 to the session No. 12001 with the pageviews (/news/default.asp, /courses/, /courses/syllabisearch.asp, /courses/syllabilist.asp, /people/facultyinfo.asp?id=231) each with a recommendation score. That's mean the recommendation systems has chooses the profile No. 5 (programs and advising) as the best source for predicting the next visited pages since this session contains pages pertaining courses.

Figure 5 shows the mean precision accuracy values obtained for different experiments conducted with

Table 2: Example of recommendation set for session 12001

Session pageviews	Recommend profile	Recommendation set	Recommend score
/news/default.asp		/programs/	0.803
/courses/	5		
/courses/syllabisearch.asp		/authenticate/login.asp?section=mycti&title=mycti&urlahead	0.707
/courses/syllabilist.asp,		=studentprofile/studentprofile	
/people/facultyinfo.asp?id=231		cti/advising/display.asp?page=intranetnews	0.555

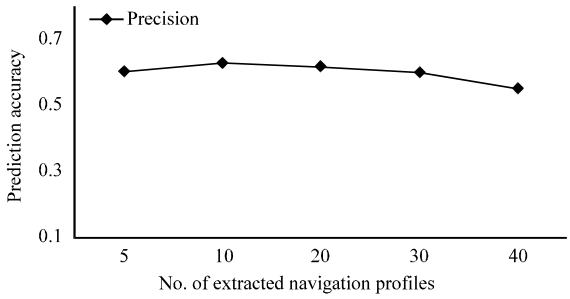

Fig. 5: Prediction Accuracy for Different Number of profiles

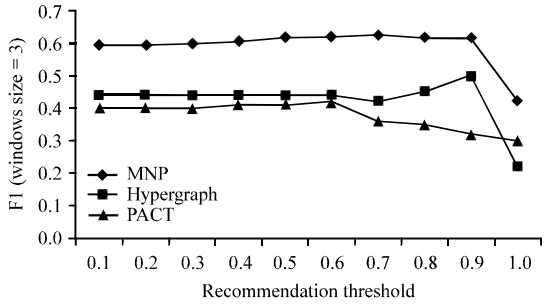

Fig. 6: Comparison of recommendation effectiveness

different number of built profiles. It shows that different number of built profiles gives different precision. This is true because the clustering size (number of built profiles in return) does matter for calculating the confidence support and the mean weights. Since, the test file includes 25% of the dataset, the precision, coverage and F1 for each transaction t is computed. Then, the final precision, coverage and F1 will be the mean precision, coverage and F1 of all transactions.

DISCUSSION

We conclude by giving some observations based on the above experimental results. The aim of the study was to evaluate the ability on predicting the new visiting page to the current active navigation session for any unidentified user in their navigation early stages using CTI dataset for the testing purpose. Based on the F1 evaluation measure, the experimental results support that anonymous usage data techniques showed promise in creating effective personalization solutions that can help attract and direct unidentified visitors based on their navigations in the early stages of their visits. Mobasher *et al.* (2002) suggested that such personalization technique gives an advantage of usage-based web personalization over traditional collaborative filtering techniques which must rely on deeper knowledge of users or on subjective input from users (such as movies or books ratings). Figure 6 relates the recommendation effectiveness for our system (Mined Navigation Profiles MNP) compared to the findings of two previous methods namely, Hypergraph and PACT (Mobasher *et al.*, 2002) reimplementation with sliding window equal to 3 using CTI dataset. With a recommendation threshold varies from 0.1 to 1.0, the F1 measurement as a performance evaluation shows that our system performs better and achieves higher prediction accuracy. This improvement is due to the mining processes applied to the extracted navigation profiles by the offline component. The mining process works on three steps: first, preprocessing the extracted navigation profiles. Secondly, enhanced classification of the active session to the best profiles and makes a recommendation list. Finally, only recommended pages with recommendation scores higher than the recommendation threshed are introduced to the user. This process ensures that the online component correctly classified the active sessions to the best mined navigation profiles. We used CTI dataset(cti.tra file) which contains only the navigations sessions without any users identifications to be distinguished from others. This file in combination with sliding Window *n* helps to test the prediction accuracy for short term navigation users. Our final observation is that the navigation profiles should be extracted again due to the navigation of many users and the change of their login time or interests which is a time consuming. Extracting incremental and adaptive navigation profiles will be more suitable for the prediction.

CONCLUSION AND FUTURE WORKS

Web is one of the main sources of the information. The ability of predicting the next visited pages and recommending it to the short term navigation user (unidentified user) is highly recommended specially in

e-commerce applications. Anonymous usage data gives an advantage of usage-based web personalization over traditional collaborative filtering techniques which must rely on deeper knowledge of users or on subjective input from users. We improved the ability to recommend to the user by learning and mining the navigation profiles for similar interested users without any deep rely on a specific user navigation history. The results showed promised prediction accuracy which helps the companies to attract and direct the anonymous users at their early stages of web navigation. Rebuilding the profiles is a time consuming process. Incremental and Adaptive profiles are one of the future interests.

REFERENCES

Baraglia, R. and F. Silvestri, 2004. An online recommender system for large web sites. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Sept. 20–24, IEEE Computer Society, Washington DC, USA., pp: 199-205.

Baraglia, R. and F. Silvestri, 2007. Dynamic personalization of web sites without user intervention. Commun. ACM., 50: 67-67.

Burke, R., 2002. Hybrid recommender systems: Survey and experiments. User Model. User-Adapted Interaction, 12: 331-370.

Deshpande, M. and G. Karypis, 2004. Item-based top-n recommendation algorithms. ACM Trans. Inform. Syst., 22: 143-177.

Fayyad, U.M., G. Piatetsky-Shapiro and P. Smyth, 1996. From Data Mining to Knowledge Discovery: An Overview. In: Advances in Knowledge Discovery and Data Mining, Fayyad, U.M., G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (Eds.). AAAI/MIT Press, Menlo Park, CA., pp: 1-34.

Jalali, M., N. Mustapha, M.N.B. Sulaiman and A.A. Mamat, 2008. Web usage mining approach based on LCS algorithm in online predicting recommendation systems. Proceedings of the 12th International Conference Information on Visualization, July 09-11, UK IEEE Computer Society Washington, DC, USA., pp: 302-307.

Joachims, T., D. Freitag and T. Mitchell, 1997. WebWatcher: A tour guide for the world wide web. Proc. Int. Joint Conf. Artif. Intel., 1: 770-775 (In Japanese).

Liu, B., 2007. Web Data Mining: Exploring Hyperlinks, Contents and Usage Data (Data-Centric Systems and Applications). Illustrated Edn., Springer, New York, ISBN-10: 3540378812, pp: 532.

Liu, H. and V. Kešelj, 2007. Combined mining of web server logs and web contents for classifying user navigation patterns and predicting users' future requests. Data Knowledge Eng., 61: 304-330.

Markov, Z. and D.T. Larose, 2007. Data Mining the Web: Uncovering Patterns in Web Content, Structure and Usage. Wiley-Interscience, Hoboken, New Jersey.

Mobasher, B., H. Dai, T. Luo, Y. Sun and J. Zhu, 2000a. Integrating web usage and content mining for more effective personalization. LNCS, Vol. 1875, Proceedings of the 1st International Conference in Electronic Commerce and Web Technologies, Sept. 04-06, Springer-Verlag London, UK., pp: 165-176.

Mobasher, B., R. Cooley and J. Srivastava, 2000b. Automatic personalization based on web usage mining. Commun. ACM., 43: 142-151.

Mobasher, B., H. Dai, T. Luo and M. Nakagawa, 2002. Discovery and evaluation of aggregate usage profiles for web personalization. Data Min. Knowledge Discov., 6: 61-82.

Nakagawa, M. and B. Mobasher, 2003. A hybrid web personalization model based on site connectivity. Proceedings of the WebKDD Workshop, (KDD'03), Washington, DC, pp: 1-11.

Srivastava, J., R. Cooley, M. Deshpande and P.N. Tan, 2000. Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explorat., 1: 12-23.

Suryavanshi, B.S., N. Shiri and S.P. Mudur, 2006. Adaptive Web Usage Profiling. In: Book Advances in Web Mining and Web Usage Analysis, Nasranui *et al.* (Eds.). LNCS, 4198, Springer-Verlag, Berlin, Heidelberg, pp: 119-138.

Taghipour, N., A. Kardan and S.S. Ghidary, 2007. Usage-based web recommendations: A reinforcement learning approach. Proceedings of the 2007 ACM Conference on Recommender Systems, Oct. 19-20, ACM, New York, USA., pp: 113-120.

Yan, T.W., M. Jacobsen, H. Garcia-Molina and U. Dayal, 1996. From user access patterns to dynamic hypertext linking. Comput. Networks ISDN Syst., 28: 1007-1014.

Yang, Q. and X. Wu, 2006. 10 Challenging problems in data mining research. Int. J. Inform. Technol. Decision Making, 5: 597-604.