International Journal of Computer Engineering Research Vol. 2(4), pp. 68-74, October 2011
Available online at http://www.academicjournals.org/IJCER
ISSN 2141- 6494 ©2011 Academic Journals

Full Length Research Paper

A least square approach to analyze usage data for effective web personalization

S. S. Patil

Rajarambapu Institute of Technology Rajaramnagar/CSE, Sangli, India. E-mail: patil.sachin.s@gmail.com.

Accepted 9 September, 2011

Web server logs have abundant information about the nature of users accessing it. The analysis of the user's current interest based on the navigational behavior may help the organizations to guide the users in their browsing activity and obtain relevant information in a shorter span of time (Sumathi and Padmaja, 2010). Web usage mining is used to discover interesting user navigation patterns and can be applied to many real-world problems, such as improving Web sites/pages, making additional topic or product recommendations, user/customer behavior studies, etc (Ratanakumar, 2010). Web usage mining, in conjunction with standard approaches to personalization helps to address some of the shortcomings of these techniques, including reliance on subjective lack of scalability, poor performance, user ratings and sparse data (Mobasher et al., 2002; Eirinaki and Vazirgiannis, 2003; Khalil et al., 2008; Forsati et al., 2009; Mobasher et al., 2001). But, it is not sufficient to discover patterns from usage data for performing the personalization tasks. It is necessary to derive a good quality of aggregate usage profiles which indeed will help to devise efficient recommendation for web personalization (Cooley et al., 1997; Srivatsava et al., 2000; Agarwal and Srikant, 1994). Also, the unsupervised and competitive learning algorithms has help to efficiently cluster user based access patterns by mining web logs (Hartigan and Wong, 1979; Ng et al., 2007; Memon and Dagli, 2003). This paper presents and experimentally evaluates a technique for finely tuning user clusters based on similar web access patterns on their usage profiles by approximating through least square approach. Each cluster is having users with similar browsing patterns. These clusters are useful in web personalization so that it communicates better with its users. Experimental results indicate that using the generated aggregate usage profiles with approximating clusters through least square approach effectively personalize at early stages of user visits to a site without deeper knowledge about them.

Key words: Aggregate usage profile, least square approach, web personalization, recommender systems, expectation maximization.

INTRODUCTION

Tremendous growth of unstructured information available on internet and e-commerce sites makes it very difficult to access relevant information quickly and efficiently. Web data research has encountered a lot of challenges such as scalability, multimedia and temporal issues etc. Web user drowns to huge information facing the problem of overloaded information.

Web personalization is the process based on users past behavior for providing users with relevant content including massive information of web pages link, relevant data, products etc. Traditionally, collaborative filtering technique was employed to do this task. It generally produces recommendations on objects yet not rated by user, by matching the ratings of current user for objects with those of similar users. To increase the user click rate and service quality of Internet on a specific website, Web developer or designer needs to know what the user really wants to do and its interest to customize web pages to the user by learning its navigational pattern. Various approaches are defined to unreal the applicative techniques to get higher and corrective recommendations for user surf.

OVERVIEW OF RELATED WORK

Various approaches have been devised for recommender systems (Mobasher et al., 2002; Eirinaki and Vazirgiannis, 2003; Khalil et al., 2008; Forsati et al., 2009). The explicit feedback from the user or rating on items help to match interest with online clustering of users with "similar interest" to provide recommendations. But practically, it leads toward limitations of scalability and performance (Mobasher et al., 2001) due to the lack of sufficient user information. Other approaches relating usage mining are implied to discover patterns or usage profiles from implicit feedback such as page visits of users.

The offline pattern discovery using numerous data mining techniques are used to provide dynamic recommendations based on the user's short term interest. "Web Personalizer" a usage based Web Personalization system using Web mining techniques to provide dynamic recommendations was proposed by Mobasher et al. (2001). According to Mehrdad et al. (2009), a novel approach using LCS algorithm improves the quality of the recommendations system for predictions by classifying user navigation patterns. K-means clustering followed by classification for recommender systems (AlMurtadha et al., 2010) is used to predict the future navigations and has improved the accuracy of predictions. Recent developments for online personalization through usage mining have been proposed. In Mobasher et al. (2002), experimental evaluation of two different techniques such as PACT and ARHP based on the clustering of user transactions/pageviews, respectively for the discovery of usage profiles was proposed.

According to Şule and Ozsu (2003), Poisson parameters to determine the recommendation scores helped to focus on the discovery of user's interest in a session using clustering approach. They are used to recommend pages to the user. This novel approach in Şule and Ozsu (2003) involving integrated clustering, association rules and Markov models improved web page prediction accuracy. Various clustering algorithms had helped to group the user sessions as like K-means, Fuzzy C-means and subtractive clustering (Chiu, 1994; Bezdek, 1973; Ratanakumar, 2010; Memon and Dagli, 2003). The clusters formed as a result of applying these algorithms are aggregated to form web profiles. The recommendation engine uses these profiles, to generate pages for recommendation. In Vasumathi and Govardhan (2005) and Spiliopoulou (2000), formal concept analysis approach is used to discover user access patterns represented as association rules from web logs which can then be used for personalization and recommendation.

However, the existing system does not satisfy users particularly in large web sites in terms of the quality of recommendations. This paper proposes to classify user navigation patterns through web usage mining system and effectively provide online recommendation. The tested results on ritindia.edu dataset indicate to improve the quality of the system for recommendations.

METHODOLOGY

Personalization using usage mining consists of four basic stages. The process embeds of:

1. Data Preprocessing
2. Pattern discovery
3. Pattern recognition
4. Recommendation process

Classifying and matching an online user based on his browsing interests for recommendations of unvisited pages has been employed in this paper using usage mining to determine the interest of "similar" users. The recommendation (Mobasher et al., 2000) consists of offline component and online component. The offline component involves data preprocessing, pattern discovery and pattern analysis. The outcome of the offline component is the derivation of aggregate usage profiles using web usage mining techniques. The online component is responsible for matching the current user's profile to the aggregate usage profiles to generate the necessary recommendations.

Data preprocessing

Preprocessing is the primary task of personalization involving cleansing of data, session and user identification, page view and transaction identification (Mobasher et al., 2000; Sumathi et al., 2010; Suresh and Padmajavalli, 2006). Let there be set of pages P = $\{p_1,p_2,p_3,p_4,\ldots.,p_n\}$ and set of n sessions, S = $\{ s_1,s_2,s_{3,}\ldots..,s_n\}$ where each s_i є S is a subset of P. A file consisting of session profile of user requests for pages is maintained.

In Sumathi and Padmaja (2010), for a particular session, a session-pageview matrix is maintained consisting of a sequence of page requests in that session. A row representing a session and every column represents a frequency of occurrence of pageview visit in a session. Then the weight of the pageview is determined by evaluating the importance of a page in terms of the ratio of the frequency of visits to the page with respect to the overall page visits in a session and is represented by a weighted session-pageview matrix. Each session s_i is modeled as a vector over the n-dimensional space of pageviews.

Pattern discovery

The primary task of pattern discovery is to find out the hidden patterns using various mining techniques such as clustering, association rule, classification etc., which helps to uncover the user behavior with respect to the site. It is an offline task which helps to determine sessions with similar navigational patterns/interest from the user session file. Clustering technique is employed to determine the session clusters using model-based expectation maximization as in Sumathi and Padmaja (2010). The profile interest is learnt by determining an aggregate usage profile using the formula:

$$wt(pg,up_c) = 1/nc \sum_{s\epsilon c} \mathrm{w}_{\mathrm{pg}}^{s} \qquad (1)$$

Wherein $\mathrm{w}_{\mathrm{pg}}^{s}$ represents the weight of the page in session s є c and nc represents the number of sessions in cluster c. Table 1 shows the aggregate usage profiles for 6 clusters under 13 distinct categories of pageviews URLs (explained in experimental evaluation).

Pattern recognition

The individual profile effectiveness is measured using weighted average visit percentage. It is to represent the significance of user's

Table 1. Aggregate usage profiles.

Page view	C					
	C0	C1	C2	C3	C4	C5
Aboutkes	0.013	0.000	0.006	0.689	0.003	0.066
Aboutrit	0.002	0.006	0.589	0.456	0.082	0.000
Courses	0.023	0.000	0.043	0.004	0.823	0.078
Departments	0.000	0.009	0.003	0.001	0.023	0.889
Facilities	0.003	0.432	0.028	0.033	0.000	0.083
Faculty	0.007	0.003	0.789	0.073	0.032	0.071
Admission	0.946	0.001	0.016	0.000	0.646	0.946
Placement	0.004	0.843	0.000	0.023	0.014	0.000
Lifeatrit	0.000	0.009	0.022	0.012	0.765	0.004
Contact	0.005	0.017	0.013	0.000	0.001	0.923
Mission-vision	0.478	0.047	0.003	0.004	0.008	0.003
Achievements	0.001	0.011	0.034	0.003	0.013	0.801
Academics-at-rit	0.002	0.708	0.049	0.028	0.022	0.000

interest in the cluster. If the aggregate usage profiles consist of m clusters and k pages, then the significance in the cluster can be determined as follows:

$$\max_{\mathrm{i}=0}^{\mathrm{m}} (wt(pg_j, up_i)), 1 \leq j \leq 13 = M_{I(j)} \qquad (2)$$

Where I (j) is the index of the maximum value in each page and $M_{I(j)}$ represents the maximum value. Also the weight of page is considered as per pageview j. This maximization function is used to recommend pages to users belonging to a profile/cluster.

Recommendation process

The recommendation engine is the online component of a usage-based personalization system. The goal of personalization based on anonymous Web usage data is to compute a recommendation set for the current (active) user session, consisting of the objects (links, ads, text, products, etc.) that most closely match the current user profile. Recommendation set can represent a long/short term view of user's navigational history based on the capability to track users across visits.

The test data of user sessions are taken as sequence of pages in time order. An active window size is fixed and those many pages are taken from the user session as active session. Then the similarity between the active session and all the cluster profiles is calculated using a vector similarity measure and the most similar profile selected for recommendations (Sumathi and Padmaja, 2010; Mobasher et al., 2002; Mobasher et al., 1999). If an active session is represented as s_i and cluster as c_k, then their similarity can be measured as follows:

$$sim(\mathrm{s_j}, \mathrm{c_k}) = \frac{\sum_{i-1}^{n} w_{i,j} * w_{i,k}}{\sqrt{\sum_{i-1}^{n} w_{i,j}^2 * \sqrt{\sum_{i-1}^{n} w_{i,k}^2}}} \qquad (3)$$

Where $w_{i,j}$, represents weight of page i in active session j and $w_{i,k}$, represents weight of page i in cluster k. The method of *least squares* assumes that the best-fit similarity of a given type is the matching score that has the minimal sum of the deviations squared (*least square error*) from a given set of data. Suppose that the data points are sim (s_1,c_1), sim (s_2,c_2),...., sim (s_j,c_k) where s_j is the independent variable and c_k is the dependent variable. The fitting score (s_j) has the deviation (error) d from each data point, that is, $d_1 = c_1 - (s_1)$, $d_2 = c_2 - (s_2)$... $d_n = c_n - (s_n)$. According to the method of *least squares*, the best matching score has the property that:

$$\Pi = d_1^2 + d_2^2 + \ldots + d_n^2 = \sum_{i-1}^{n} d_i^2 = \sum_{i-1}^{n} [c_i - \mathrm{f}(s_i)]^2 = \text{a min.} \qquad (4)$$

$$\text{sim'(sj,ck)} = \text{sim (sj,ck)} - \Pi \qquad (5)$$

This *least square error* approach helps to fine tune the scores such as to approximate user clusters based on similar web access patterns on their usage profiles. Then a recommendation score for each page view p in the selected cluster/profile is calculated. If C is the most similar cluster/profile to the active session S, then a recommendation score for each page view p in C is as:

$$\mathrm{Rec(S, p)} = \sqrt{\mathrm{weight(p, C)} * \mathrm{sim}'(}\mathrm{s_j, c_k}) \qquad (6)$$

Profiles having a similarity greater than a threshold value μ_c are selected as matching clusters in the decreasing order of their scores. The weight of pageview p in C is computed twice that is, directly and indirectly but to compensate the impact, square root in the above function is taken and results are normalized to value between 0 and 1. If the pageview p is in the current active session, then its recommendation value is set to zero. These matching clusters can be used for recommending pages instantaneously which have not been visited by the user. Figure 1 shows the overall process of web personalization using web usage data.

Experimental evaluation

This section provides a detailed experimental evaluation of the profile generation techniques. The privately available data set at the University of Shivaji, containing web log files of ritindia.edu web site have been used for this research. It includes the page visits of users who visited the "ritindia.edu" web site in period of June 2006 to April 2011. The initial log file produced a total of 16,233 transactions and the total number of URLs representing pageviews was 27. By using support filtering for long transactions, pageviews appearing in less than 0.5% or more than 85% of transactions were eliminated. Also, short transactions with at least 5 references were eliminated. The visits are recorded at the level of URL category and in time order, which includes visits to major 13 distinct categories of pageviews URLs. Each sequence in the dataset corresponds to a user's request for a page. The 13 categories are shown in Table 2.

A clustering model is estimated using approximately 15,000 samples within the dataset. They are further classified into training and testing sets. Dataset is split into 70% training and 30% testing sets such that the model is designed using training set and then evaluated using test samples for performance. Applying the clustering algorithm for Expectation Minimization with 12 iterations

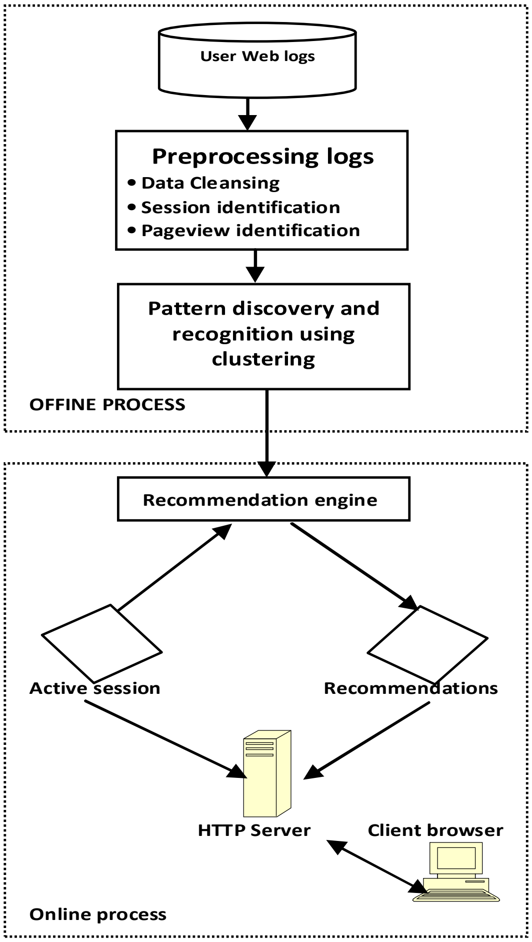

Figure 1. The overall process of personalization.

Table 2. Thirteen distinct categories of page views URLs.

Id	Category
1	aboutkes
2	aboutrit
3	courses
4	departments
5	facilities
6	faculty
7	admission
8	placement
9	lifeatrit
10	contact
11	mission-vision
12	achievements
13	academics-at- rit

results in 6 major clusters. Each cluster represents several sessions of navigational patterns representing “similar” interest in the web pages or the usage profile and the aggregate usage profile is determined using Equation 1. During the online phase, the pages visited in a session are stored in a user session file and after each page visit, the relative frequency of pageviews in the active session is determined. An active session with sliding window size ‘n’ (in our experiment, the size is 5 as it represents the average number of page visits in the dataset) consists of the current page visit and the most recent n-1 pages visited. The window slides, as the user

Table 3. Page visits in the sliding window.

Session	Order of visit	Window	Active session (page visited)				
1	1	1	1	0	0	0	0
1	2	2	1	8	0	0	0
2	1	1	1	0	0	0	0
2	2	2	1	8	0	0	0
2	3	3	1	8	13	0	0
3	1	1	1	0	0	0	0
3	2	2	1	13	0	0	0
3	3	3	1	13	1	0	0
3	4	4	1	13	1	13	0
3	5	5	1	13	1	13	1
3	6	6	13	1	13	1	7
3	7	7	1	13	1	7	1

Table 4. Frequency of page visited/weighted page view

Session	Order of visit	Aboutkes	Aboutrit	Courses	Department	Facilities	Faculty	Admission	Placement	Lifeatrit	Contact	Mission-vision	Achievement	Academis-at-rit
1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	2	1	0	0	0	0	0	0	1	0	0	0	0	0
2	1	1	0	0	0	0	0	0	0	0	0	0	0	0
2	2	1	0	0	0	0	0	0	1	0	0	0	0	0
2	3	1	0	0	0	0	0	0	1	0	0	0	0	1
3	1	1	0	0	0	0	0	0	0	0	0	0	0	0
3	2	1	0	0	0	0	0	0	0	0	0	0	0	1
3	3	2	0	0	0	0	0	0	0	0	0	0	0	1
3	4	2	0	0	0	0	0	0	0	0	0	0	0	2
3	5	3	0	0	0	0	0	0	0	0	0	0	0	2
3	6	2	0	0	0	0	0	1	0	0	0	0	0	2
3	7	3	0	0	0	0	0	1	0	0	0	0	0	1

browses through various pages. Now, using the cosine similarity measure, the active session is matched with the aggregate usage profiles and matching cluster(s) having value greater than the threshold and are used for recommending pages exceeding threshold that have not been visited by the user. For example, consider the following 3 sessions consisting of page visits:

1 8
1 8 13
1 13 1 13 1 7 1

The sliding window consists of the pages 1 13 1 13 1 in the fifth page visit. Tables 3 and 4 represent the page visits and frequency of visited pages in the sliding window.

Table 5. Matching clusters.

Session	Order of visit/ window	Page visited	C0	C1	C2	C3	C4	C5
1	1	1	0.000	0.000	0.245	0.000	0.169	0.000
1	2	8	0.000	0.000	0.000	0.076	0.243	0.000
2	1	1	0.023	0.000	0.334	0.000	0.043	0.000
2	2	8	0.000	0.000	0.332	0.031	0.223	0.013
2	3	13	0.025	0.192	0.000	0.000	0.332	0.000
3	1	1	0.000	0.023	0.034	0.333	0.000	0.000
3	2	13	0.020	0.000	0.000	0.014	0.223	0.336
3	3	1	0.000	0.012	0.000	0.443	0.000	0.020
3	4	13	0.013	0.000	0.043	0.258	0.014	0.256
3	5	1	0.000	0.012	0.033	0.000	0.000	0.276
3	6	7	0.000	0.013	0.344	0.000	0.017	0.000
3	7	1	0.012	0.000	0.000	0.344	0.000	0.032

Table 6. Recommendation set for session 1, session 2 and session 3.

Session	Order of visit/ window	Active session window (pages)	Matching cluster(s)	Recommendation pages
1	1	1	2,4	2,3,6,7,9
1	2	1->8	2,3,4	2,3,6,7,9
2	1	1	2,4	2,3,6,7,9
2	2	1->8	2,3,4	2,3,6,7,9
2	3	1->8->13	1,2,3,4	2,3,6,7,9,11
3	1	1	2,4	2,3,6,7,9
3	2	1->13	2,4,5	2,3,4,6,7,9, 10,12
3	3	1->13->1	2,3,4,5	2,3,4,6,7,9, 10,12
3	4	1->13->1->13	2,3,4,5	2,3,4,6,7,9, 10,12
3	5	1->13->1->13->1	2,3,4,5	2,3,4,6,7,9, 10,12
3	6	13->1->13->1->7	2,3,4,5	2,3,4,6,9,10,12
3	7	1->13->1->7->1	2,3,4,5	2,3,4,6,9,10,12

It states the weight of the pageview by evaluating the importance of a page in terms of the ratio of the frequency of visits to the page with respect to the overall page visits in the active session.

Clusters greater than the threshold value, are chosen to be matching clusters as shown in Table5. This table depicts the comparative study of aggregate usage profiles and the maximization function to show recommendations. It has been found that when the user visits page 1 (window size 1), the appropriate clusters, exceeding the threshold value are cluster 2 and 4. It is seen that pages 2, 3, 6, 7, 9 can be recommended from cluster 2 and 4. Similarly, when the user visits page 8 subsequent to page visit 1(window size 2), the appropriate matching clusters are cluster 2, cluster 3 and cluster 4. As the window size increases to the fixed size limit (n = 5), correspondingly, the matching clusters for the visited page(s) in the active session and the recommendations are dynamic in nature. Table 6 shows the recommended set of pages for all 3 demo sessions.

As compared to experimental study in Sumathi and Padmaja (2010), the recommendation of pages for demo sessions is calculated on Equation 3 for measuring similarities of sessions. Then the pages are recommended in the session as per Equation 6 for simple similarities of clusters which provides a less precision for recommending the matching scores of clusters. As compared to same, the least square approach analysis in our study helps to find deviations and to refine further the recommendations as per Equation 6 with similarity measures. In Sumathi and Padmaja (2010), if the previous 3 demo samples are measured, then the recommendation set varies a lot recommending less pages under given threshold.

CONCLUSION AND FUTURE WORK

The ability to collect detailed usage data at the level of individual mouse click provides Web-based companies with a tremendous opportunity for personalizing the Web experience of clients. The practicality of employing Web usage mining techniques for personalization is directly related to the discovery of effective aggregate profiles that can successfully capture relevant user navigational patterns and can be used as part of usage-based recommender system to provide real-time personalization.

In this work, the primary objective was to classify and match an online user based on his browsing interests.

Identification of the current interests of the user based on the short-term navigational patterns instead of explicit user information has proved to be one of the potential sources for recommendation of pages. In particular context of anonymous usage data, these work under least square approximation show promise in creating effective personalization solutions that can help retain and convert unidentified visitors based on their activities in the early stages of their visits. Future work involves with various types of transactions derived from user sessions, such as to isolate specific types of "content" pages in the recommendation process. Also, the plan is to incorporate client-side agents and use of optimization techniques to assess the quality of recommendations.

REFERENCES

Agarwal R, Srikant R (1994). Fast Algorithms for Mining Association Rules in Large Database. Procd. Conf. on Very Large Data Bases, USA., 1: 12-15.

AlMurtadha Y, Sulaiman M, Mustapha N, Udzir N (2010). Mining Web Navigation Profiles for Recommendation System. Inform. Technol. J., 9: 790-796.

Bezdek J (1973). Fuzzy Mathematics in Pattern Classification, Cornell Univ. PhD Thesis.

Chiu S (1994). Fuzzy Model Identification Based on Cluster Estimation. J. Intelligent Fuzzy Syst., 2(3): 268-278

Cooley R, Mobasher B, Srivatsava J (1997). Web Mining: Information and Pattern Discovery on the World Wide Web. IEEE., 9: 558-567.

Eirinaki M, Vazirgiannis M (2003). Web mining for Web personalization.ACM, 3(1): 1-27

Forsati R, Meybodi M, Ghari N (2009). Web Page Personalization based on Weighted Association Rules. Internet Conf. Electron. Comp.Tech., 1: 130-135.

Hartigan T, Wong A (1979). A K-Means Clustering Algorithm. Appl. Statistics, 28: 100-108.

Khalil F, Jiuyong L,Hua W (2008). Integrating Recommendation Models for Improved Web Page Prediction Accuracy. Australa. Conf. Comp. Sci., 74: 91-100.

Mehrdad J, Norwati M, Ali M, Nasir B (2009). A Recommender System for Online Personalization in the WUM Applications. Procd. World Congress Engr. Comp. Sci., 9(2).

Memon K, Dagli C (2003). Web Personalization using Neuro-Fuzzy Clustering. IEEE., 1: 525-529.

Mobasher B, Cooley R, Srivatsava J (1999). Creating Adaptive Web Sites Through Usage-Based Clustering of URLs. Proc. KDEX '99. Workshop on Knowledge and Data Engineering Exchange.

Mobasher B, Cooley R, Srivatsava J (2000). Automatic Personalization Based on Web Usage Mining. ACM., 43:142-151.

Mobasher B, Dai H, Luoand T, Nakagawa (2001). Improving the effectiveness of collaborative filtering on anonymous Web usage data. IJCAI, Seattle.

Mobasher B, Honghua D, Tao L, Nakagawa M (2002). Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization. ACM., 6: 61-82.

Ng K, Junjie M, Huang J, Zengyou Y (2007). On the Impact of Dissimilarities Measure in K-modes Clustering Algo. IEEE., 29(3).

Ratanakumar J (2010). An Implementation of Web Personalization Using Mining Techniques. J. Theoretical Appl. I. T., 5(1): 67-73.

Spiliopoulou M (2000). Web usage mining for Web Site Evaluation. ACM., 43(8): 127-134.

Srivatsava J, Cooley R, Deshpande M, Tan P (2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. ACM., 1: 12-23.

Şule G, Ozsu M (2003). A User Interest Model for Web Page Navigation. DMAK, Seoul, 1: 46-57.

Sumathi C, Padmaja R, Valli, Santhanam T (2010). J. Comp. Sci., 1: 785-793.

Sumathi CP, Padmaja Valli R (2010).Automatic Recommendation of Web Pages in Web Usage Mining. Internet J. Comp. Sci. Engr., 2(9).

Suresh R, Padmajavalli R (2006). Overview of Data Preprocessing in Data and Web Usage Mining. IEEE., 1: 193-198.

Vasumathi D, Govardhan A (2005). Efficient Web Usage Mining Based On Formal Concept Analysis. IFIP., 163(2): 99-108.