AENSI Journals

Australian Journal of Basic and Applied Sciences

ISSN:1991-8178

Journal home page: www.ajbasweb.com

Discovering user profiles for web personalization using EM with Bayesian Classification

[1]T.Gopalakrishnan and [2]Dr. P. Sengottvelan

[1]*Assistant Professor, Department of Information Technology, Bannari Amman Institute of Technology Sathyamangalam, India.*
[2]*Associate Professor, Department of Information Technology, , Bannari Amman Institute of Technology, Sathyamangalam, India.*

ARTICLE INFO

Article history:
Received 25 January 2014
Received in revised form 8 March 2014
Accepted 10 March 2014
Available online 3 April 2014

Keywords:
Web Personalization, User Profile, Weighted EM algorithm with naive bayseian, Clustering

ABSTRACT

Background: With the continued increase and explosion of Web-based applications, Web services, the volumes of click-stream and user data, collected by respected enterprises in their daily operations have reached excessive peak amount. The analysis of the user's current interest based on their navigational behavior may help the organizations to guide their users in their browsing activity and gain appropriate information in a shorter period of time. This type of analysis is used for the automatic discovery of users interest on the specific web pages which are often stored in web server access logs. ***Objective:*** This paper focuses on recommending similar web pages by tracking the usage behavior of a web user at various sessions and clustering them with similar interest user's through expectation maximization technique with naïve Bayesian classifier. This technique also considers time spent by the user on the specific page and aggregates user profiles. ***Results:*** Experimental setup results in the creation of usage profiles, identification of user interest and effective recommendation of similar web pages by analyzing navigational behavior of the user than the other traditional methods. ***Conclusion:*** The results of the implemented methods will help in improving the performance of Web information retrieval based on their navigational interest and automatic recommendations for the web sites yet to be visited by the user.

To Cite This Article: T. Gopalakrishnan, Dr. P. Sengottvelan, Discovering user profiles for web personalization using EM with Bayesian Classification, ***Aust. J. Basic & Appl. Sci.,*** *8(3): 53-60, 2014*

INTRODUCTION

Day-to-day world faces constant increase of Web-based applications systems, information systems, click-stream data's and user data are collected by Web-based organizations for their every day operations almost in all the fields. The extraction of the hidden pattern and creating an useful information is a most important ability in the present day web environment. The web server-logs and click-stream data will be helpful to model and analyze the users' browsing behaviour pattern. The web-server logs mostly have partially structured data and analyzing these stored log data would provide the automatic discovery of the useful information. The automatic discovery of usage profiles from captured click-stream and user data will allow the organizations to improve their web site management and provide personalized recommendations of web pages according to the current interests of the user.

Web Mining can be specified as the discovery and analysis of useful information from the World Wide Web (cp. Cooley R., 1997). Web Usage Mining attempts to make sense of the data generated by the Web surfer's session data or behaviours patterns. with the intention to understand and serve the needs of Web applications, Usage Data captures the identity and origin of Web users along with their browsing behaviour at a Web site. This data includes data from registration data, user profiles, Web server access, proxy server and browser logs, user sessions or transactions, cookies, user queries, mouse clicks and any other data that results from interactions.

Corresponding Author: T. Gopalakrishnan, Department of Information Technology, Bannari Amman Institute of Technology, Sathyamagalam. India.
E-mail: gopalakrishnan.ct@gmail.com

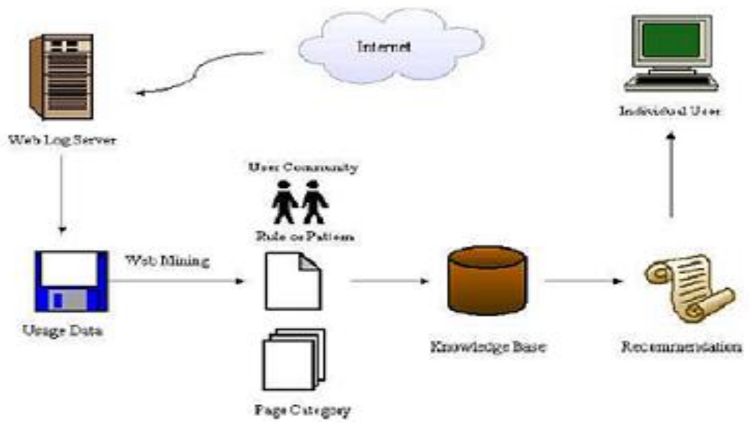

Fig. 1: Steps involved in web personalization

In web personalization the knowledge discovery part can be executed *offline* and periodical mining provides new contents of the user access log files, and automatic recommendation can be done in the following steps:

(a) ***collection** of raw data* from the log server,
(b) modelling and categorization of these data **(*pre-processing*** *phase*),
(c**)** ***analysis** of the user behaviour* from the collected data(Pattern Discovery and Analysis)
(d) suggesting of web pages to the individual user(***Recommendation*** phase)

The ways that are employed in order to analyze the collected data include content-based filtering, collaborative filtering, rule-based filtering, and Web usage mining. The site is personalized through the importance of existing hyperlinks, the dynamic insertion of new hyperlinks that appear to be of interest for the current user, or even the formation of new index pages.

This paper focuses on the automatic discovery of “similar” interests of groups of sessions based on the user's navigation behaviour. To attain this:

- Intially, a usage model is developed based on the browsing behaviour during various sessions
- Using this model, we create “similar” interests of groups of sessions ,called as aggregate usage profile
- Each usage profile consists of pages with varying user interest/significance
- In this proposed method, the weight of the pages in each profile is determined. These profiles would later help in various applications of web usage mining such as assigning a new user to the appropriate cluster, web site management, and recommend pages of interest yet to be visited by the user to offer personalized web content.

Related Work:

Lot of research has focused on the web usage mining algorithms for mining user navigation behaviour and patterns. In this following we will review some of the major navigation pattern mining systems and algorithm s in web usage mining area that can be compared with our system.

The partitioning method was the best clustering methods to be used in Web usage mining by Yan T W*et al*.(1996). They used an iterative algorithm that produces high quality clusters. Each user session is modelled by an *n*-dimensional vector, where *n* is the number of Web pages in the session. The value of each feature is a specified weight, measuring the degree of interest of the user in the particular Web page. This calculation is based on a number of parameters, like the number of times the particular page has been accessed and the amount of time the user spent on the page.The characterized sessions are the patterns discovered by the algorithm. The main problem with this approach is the calculation of the feature weights. The choice of the right parameter mix for the calculation of these weights is not straightforward and depends on the modelling abilities of a human expert.

There are various distance-based similarity measures such as, Manhattan distance, Mutual Neighbour Distance (MND), Simple Matching Coefficient Euclidean distance measure, Jaccard Coefficient and Rao’s coefficient. The usage of these similarity measures depends on the features of the different samples. Anyway, in Chaofeng. L (2009), a Sequence Alignment Method has used for measuring similarities between web pages by considering the URL and the viewing time of the URL. Decrease in time and space complexity has been proved in the proposed algorithm for Web Session Clustering Based on Increase of Similarities (WSCBIS) and Robust Clustering using links (ROCK).

Cadez *et al*. (2003) in the Web CANVAS proposed a partitioning clustering method, which visualize user navigation paths in each cluster. In this method, user sessions are represented using categories of general topics for Web pages. A more number of predefined categories are used as a bias, and URLs from the respective Web server log files are assigned to them, by constructing the user sessions.

In Lee and Fu, (2008), two levels of prediction of users' browsing behaviour have been proposed. Using Markov Model, browsing behaviour is predicted at the category level and using Bayes Theorem, prediction is done at the web page level. A combination of Markov model and Bayes theorem results in a two-level prediction of user's browsing behaviour. This results proved that the hit ratio is effective and accurate in both the levels. The study concerning clustering of URLs using Sequence Alignment Method has also been done in Hay B (2008) .In this study, clustered web users using two different similarity measures: SAM (non-Euclidean distance-based measure) and Association measure (Euclidean distance-based measure).

The Expectation-Maximization (EM) algorithm, Dempster A P (1977), based on mixtures of Markov chains is used for clustering user sessions. Each Markov chain represents the behaviour of a particular subgroup. EM is a memory efficient and easy to implement algorithm, with a profound probabilistic background. However, there are cases where it has a very slow linear convergence and may therefore become computationally expensive, although in the results in Cadez *et al*.(2003), it is shown empirically that the algorithm scales linearly in all aspects of the problem.

More researchers have worked to have improved quality of clustering models. These works attempt to find architecture and algorithm for categorizing the user behaviour profiling to further enhance the quality of the recommendations, but the quality still does not meet satisfaction. We propose a new modified Expectation Maximization clustering with Naive Bayesian model that handles the excess memory requirements in case of large data sets by reducing the number of paths during the training and testing phases. In addition, the results indicate a remarkable enhancement in prediction time for our novel proposed two-tier(EM- Naïve Bayesian framework).

System Design and Methodology:

Data Pre-Processing:

The data accessed in web server log file is incomplete and not fit for mining directly. Pre-processing step is necessary to convert the data into suitable form to have a pattern discovery process. Pre-processing might provide accurate, crisp data for data mining. Data pre-processing, includes data cleaning, user identification, path completion, user sessions identification, and data integration.

Let there be set of pages
Pg= $\{p_{g1},p_{g2},p_{g3},p_{g4},\ldots,p_{gn}\}$ and
set of Q sessions,

Sn= $\{ s_{n1},s_{n2},s_{n3},\ldots,s_{nn}\}$ where each $s_{ni} \in S_n$ is a subset of P.

A log consisting of session profile of user requests for pages is stored, stored and maintained at regular intervals.

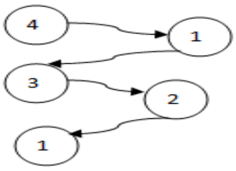

Fig. 1: Node Representation

If suppose, user access the page in following sequences 41321 represents a session consisting of a sequence of page requests the above session pageview matrix is obtained.

Table 1: Pageview-individual session wise

Session- PV Nodes	I	II	III	IV
1	4	3	1	3
2	1	2	0	1
3	3	1	2	3
4	2	2	3	4
5	1	4	4	2

Each page is mapped with a weight representing its impact and importance . The weight of the web page can be can be determined in various ways depending on the type of analysis. In most Web Usage Mining tasks, the weights may be based on a combination of factors such as the time that the user has spent on a page visited, number of visits to the page and size of the page. Hence the weight of a node pi is the sum of the weights of the in degree of the node pi:

$W(P_i) = \sum W(e_i)$

Weight of the page W_i= W($T_{s)}$ +W($N_{v)}$+W($S_{z)}$

Where, T_s is the time spent by the user on the particular page,
N_v represents the number of visits to the page
$S_{z,}$ represents the size of the page

In [10], 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 in a weighted session-pageview matrix. Each session s_{ni} is modelled as a vector over the n-dimensional space of pageviews.

Table 2: Weighted-Session Pageview

User session	Visit Order	Home Page	Events	placement	Faculty	contacts	webmail	R&D	achievements	Clubs& activites	Academics -UG	Academics -PG
1	1	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	2	0.500	0.000	0.000	0.000	0.000	0.000	0.500	0.000	0.000	0.000	0.000
1	3	0.667	0.000	0.000	0.000	0.000	0.000	0.333	0.000	0.000	0.000	0.000
2	1	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	2	0.500	0.000	0.000	0.000	0.000	0.000	0.500	0.000	0.000	0.000	0.000
2	3	0.333	0.000	0.000	0.000	0.333	0.000	0.000	0.000	0.000	0.333	0.000
3	1	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	2	0.500	0.000	0.000	0.000	0.500	0.000	0.000	0.000	0.000	0.000	0.000
3	3	0.667	0.000	0.000	0.000	0.000	0.333	0.000	0.000	0.000	0.000	0.000
3	4	0.500	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.000	0.000	0.000
3	5	0.600	0.000	0.000	0.400	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	6	0.400	0.000	0.000	0.000	0.200	0.000	0.400	0.000	0.000	0.000	0.000
3	7	0.200	0.000	0.000	0.000	0.000	0.000	0.400	0.000	0.200	0.000	0.000

Pattern Discovery:

The main purpose of the pattern identification is to obtain sessions with “similar” profile/ interests. Mostly the clustering methods such as hierarchical clustering, partition clustering are used wherein the similarity measure is model-based, distance-based which clustering employs probability-based approach. The origin of the probability based clustering method is based on finite mixture model. Mixture models are capable to capture most complex and dynamic user behaviour. The most suitable method used for this kind of problem is Expectation Maximization Technique, which can find out the mean and standard deviation in an appropriate way.

3.3 EM Algorithm with Naïve Bayesian Classification:

One of the Labelled and Unlabelled learning techniques uses the Expectation–Maximization (EM) algorithm. EM is a accepted iterative algorithm for maximum likelihood estimation in problems with normal missing data. The EM algorithm consists of following two steps, the **Expectation step** (or **E-step**), and the **Maximization step** (or **M-step**). The E-step mainly fills in the missing data based on the current estimation of the parameters. The M-step, which maximizes the likelihood, will re-estimate the parameters. EM converges to a local minimum when the model parameters gets stabilize.

The ability of EM to work with missing data is exactly what is needed for learning from labelled and unlabeled examples. The documents in the labelled set (denoted by *L*) all have class labels. The documents in the unlabeled set (denoted by *U*) can be regarded as having missing class labels. We can use EM to estimate them based on the current model, i.e., basically to assign probabilistic class labels to each document *di* in *U*, i.e., Pr(*cj*|*di*).

After a number of iterations, all probabilities will converge at a point.

Here we will use *the naïve Bayesian* (NB) algorithm as the base algorithm, and run it iteratively. The parameters that EM estimates in this case are the probability of each word given a class and the class prior probabilities.

$$Pr(W_t|C_{j,}\cdot\Theta) = \frac{\lambda + \Sigma_{i=1}^{|D|} N_n \, Pr(C_j|d_i)}{\lambda|V| + \Sigma^{|V|}_{i=1} \Sigma^{|D|}_{i=1} N_{si} \, Pr(C_j|d_i)} \tag{1}$$

$$Pr(c_j|d_i;\Theta) = \frac{Pr(c_j|\,\Theta)\; Pr(d_i|c_j;\,\Theta)}{P_r(d_i\,|\,\Theta)} \tag{2}$$

Although it is quite involved to derive the EM algorithm with the NB classifier, it is fairly straightforward to implement and to apply the algorithm. That is, we use a NB classifier in each iteration of EM, below equation

$$p_r(c_j \mid d_i;\overline{\Theta}) = \frac{p_r(c_j \mid \overline{\Theta})\, p_r(d_i \mid c_j \overline{\Theta})}{p_r(d_i \mid \overline{\Theta})}$$

$$= \frac{p_r(c_j \mid \overline{\Theta}) \prod_{k=1}^{|d|} p_r(w_{d_{i,k}} \mid c_j \overline{\Theta})}{\sum_{r=1}^{|C|} p_r(c_j \mid \overline{\Theta}) \prod_{k=1}^{|d|} p_r(w_{d_{i,k}} \mid c_j \overline{\Theta})} \tag{3}$$

for the E-step, and Equations (1) and (2) in for b the M-step. First build a NB classifier *f* using the labelled examples in *L*. Then use *f* to classify the unlabeled examples in *U*, will more accurately to assign a probability to each class for every unlabeled example, i.e., Pr(*cj*|*di*), which takes the value in [0, 1] instead of {0, 1}. Some explanations are in order here.

Let the set of classes be *C* = {*c*1, *c*2, …, *c*|*C*|}. Each iteration of EM will assign every example *di* in *U* a probability distribution on the classes that it may belong to. That is, it assigns *di* the class probabilities of Pr(*c*1|*di*), Pr(*c*2|*di*), …, Pr(*c*|*C*||*di*). Based on the assignments of Pr(*cj*|*di*) to each document in *U*, a new NB classifier can be constructed. This new classifier can use both the labelled set *L* and the unlabeled set *U* as the examples in *U* now have probabilistic labels, Pr(*cj*|*di*). This leads to the next iteration. The process continues until the classifier parameters (Pr(*wt*|*cj*) and Pr(*cj*)) no longer change (or have minimum changes).

The EM algorithm with the NB classification was proposed for LU learning . The algorithm is given below can also be seen as a clustering method with some the initial seeds in each cluster. The class labels of the seeds indicate the class labels of the resulting clusters.

```
Algorithm EM(L, U)
Learn an initial naïve Bayesian classifier f from only the labeled set(using (1) & (2) );
  repeat
// E-Step
    for each example di in U do
      Using the current classifier f to compute Pr(cj|di). (using (1) & (2)).
    end
// M-Step
    learn a new naïve Bayesian classifier f from L □ □U by computing Pr(cj) and Pr(wt|cj). (using (3)).
  until the classifier parameters stabilize Return the classifier f from the last iteration.
```

Fig. 2: The EM algorithm with naïve Bayesian classification

3.4 Experimental Evaluation:

This part provides a comprehensive investigational evaluation for the profile creation techniques. The in-private accessible data set of our institute , containing web (IIS) log files of our institute web site have been taken for this experiment and study. This includes the counts who visited our website during the academic year 2012 to 2013.The early log file produced a whole of 13,246 transactions and the total number of URLs representing pageviews was 23. By using Support filtering method for long transactions, the respective pageviews appearing in less than 3 % or more than 80% of transactions were eliminated. And, short transactions with at least of 15 references were eliminated. The visits are recorded at the level of URL category , time order, which includes visits to major 12 distinct categories of the pageviews URLs. Any sequence in the dataset refer to a user's request for a specific page.

The web site contains the 12 category of pages with their index as listed :

Index	URL address
0	/bitsathy/Home Page
1	/bitsathy/webmail
2	/bitsathy/events
3	/bitsathy/placements
4	/bitsathy/R&D
5	/bitathy/academics
6	/bitsathy/ academics/ug
7	/bitsathy/faculty
8	/bitsathy/achievements
9	/bitsathy/clubs&activites
10	/bitsathy/acdemics/pg
11	/bitsathy/contacts

The collected log dataset is split into categories which would be involved 75% training and 25% testing sets. Applying this discussed clustering algorithm, with the 12 iterations results in the following 6 major clusters. The aim of the forming cluster is to represent the several sessions of navigational patterns indicate "similar" interest in the usage profile. During this process, pages visited in a session gets stored in a user session file and after each page visit by the user, the relative frequency of the respective pageviews in the active session is determined. The active session can be represented with sliding window size 'n' (in our test, the size is 5 as it represents the average number of page visits in the dataset) which consists of the current page visit and also the most recent n-1 pages visited. Now, Using the cosine similarity measure, the active session is matched with the calculated aggregate usage profiles and with the matching cluster(s) which are having value greater than and nearby threshold are used for recommending pages that exceeds threshold that have not been visited by the user.

To make an analysis, consider the following 3 sessions consisting of page visits.
5 6 4
5 6 10
5 10 5 10 5 7 5

Table 3: Total Page Visits

session	1	1	1	2	2	2	3	3	3	3	3	3	3
Visit order	1	2	3	1	2	3	1	2	3	4	5	6	7
Page Visited or session active	5	5	5	5	5	5	5	5	5	5	5	5	5
	0	6	6	0	6	6	0	10	10	10	10	10	10
	0	0	4	0	0	10	0	0	5	5	5	5	5
	0	0	0	0	0	0	0	0	0	10	10	10	10
	0	0	0	0	0	0	0	0	0	0	5	7	5

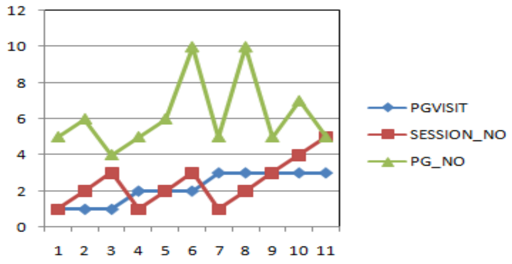

Fig. 2: session wise pagevisit

The representation of the sliding window consists of the pages 5 10 5 10 5 7 5 in the third session visit. Table 3 shown represents the page visits. Table 4 states the weight of the pageview by evaluating the significance of a page in terms of the ratio of the frequency of visits to the page with respect to the overall page visits during the current the active session.

Table 4: Matching Clusters

Count Session	User Visit order	Page Visited	C0	C1	C2	C3	C5	C5
1	1	5	0.000	0.000	0.233	0.000	0.103	0.000
1	2	6	0.000	0.000	0.089	0.000	0.240	0.155
1	3	4	0.000	0.000	0.096	0.000	0.195	0.000
2	1	5	0.000	0.000	0.078	0.000	0.214	0.000
2	2	6	0.000	0.124	0.161	0.000	0.141	0.000
2	3	10	0.000	0.000	0.056	0.000	0.191	0.000
3	1	5	0.000	0.000	0.150	0.000	0.212	0.000
3	2	10	0.000	0.000	0.075	0.000	0.144	0.121
3	3	5	0.000	0.000	0.161	0.000	0.321	0.000
3	4	10	0.000	0.000	0.140	0.078	0.165	0.000
3	5	5	0.000	0.136	0.122	0.000	0.222	0.000
3	6	7	0.000	0.000	0.135	0.000	0.187	0.000

Formed Clusters which have greater than the threshold value, are chosen to be matching clusters as shown in Table 4. This given table depicts the comparative study of aggregate usage profiles and the expectation maximization function to show the recommendation pages. It has been identified that when the user visits page 5 (window size 1), the appropriate clusters, exceeding the threshold value are cluster 2 and 5.

It is seen that pages 5, 6,10 can be recommended from cluster 3 and 5. Similarly when the user visits page 6 subsequent to page visit 1(window size 2), the appropriate matching clusters are cluster 1, cluster 4 and cluster 5. 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 5 shows the recommended set of pages for all three demo sessions.

Table 5: Page set recommended during sequence Session

Session	User's Visit Order	Pages Active session	Similar Clusters	Pages Recommended
1	1	5	3,5	5,6,10
1	2	5 ->6	1,3,5	5,6,10
1	3	5 ->6->4	1,3,5	5,6,7,10
2	1	5	1,5	5,6,10
2	2	5 ->6	1,4,5	5,6,10
2	3	5 ->6->10	1,3	5,6,7,8,10
3	1	5	5	5,6,10
3	2	5 ->10	3,5	5,6,10
3	3	5 ->10->5	3,4,5	5,6,10
3	4	5 ->10->5->10	2,5	5,6,7,8,10
3	5	5->10->5	2,4	5,6,10
3	6	5->10->5->7	4,5	5,6,7,8,10
3	7	5->10->5->7->5	2,3,4,5	5,6,7,8,10

Conclusion:

This study shows method to collect in-depth usage data, at the level of individual users provides Web-based enterprise with a tremendous chance for personalizing the Web. The possibility of employing Web usage mining techniques for personalization is the process of discovering effective aggregate profiles that can successfully used to capture relevant user navigational patterns which could be used as part of web page recommender system.

Here, the main aim is to classify and match web user based on his browsing interests. Discovery of the user's current interest is also based on the short-term navigational patterns instead of explicit collection of user's information has proved to be one of the potential sources for recommendation of pages. Initially, we mapped the frequency of page visits to the relative user interest during a session and calculate a weighted session-pageview matrix. With this calculated weight, model-based Expectation Maximization clustering algorithm is applied to categorize clusters or usage profiles. Experiment are done with real world data set to identify the user interest in the web. The importance of the search results will be useful for the enterprises for their web site improvement based on the user's navigational interest and provide them with similar recommendations for page(s) yet visited be by the user.

REFERENCES

Cadez, D., C. Heckerman, P. Meek, Smyth and S. White, 2000. Visualization of navigation patterns on a Web site using model-based clustering. Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pp: 280-284.

Chaofeng, L., 2009. Research on web session clustering. J. Software, 4: 460-468, DOI:10.4304/jsw.4.5.460-468.

Dempster, A.P., N.M. Laird and D.B. Rubin 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, (39): 1-38.

Han, J. and M. Kamber, 2006. Data Mining: Concepts and Techniques, 2nd Edn., Morgan Kaufmann Publishers, San Francisco, CA., ISBN: 978-1- 55860-901-3, pp: 770.

Hay, B., G. Wets and K. Vanhoof, 2008. Clustering navigation patterns on a website using a sequence alignment method. Knowl. Inform. Syst., 6: 150-163. DOI: 10.1007/BF02637153.

Lee, C.H. and Y.H. Fu, 2008. Two levels of prediction model for user's browsing behaviour. Proceeding of the International Multi Conference of Engineers and Computer Scientists, National Science Council, Hong Kong, pp: 751-756.

Mohammad-R. Akbarzadeh-T. 2 Noorali Raeeji Yanehsari, 2009 .Web Usage Mining: users' navigational patterns extraction from web logs using Ant-based Clustering Method, Kobra Etminani, IFSA-EUSFLAT.

Norwati Mustapha, Manijeh Jalali, Abolghasem Bozorgniya, Mehrdad Jalali, 2009. Navigation Patterns Mining Approach based on Expectation Maximization Algorithm. World Academy of Science, Engineering and Technology, (26): 2009-02-26.

Srivatsava, J., R. Cooley, M. Deshpande and PN. Tan, 2010. Web usage mining: discovery and applications of usage patterns from Web data. ACM SIGKDD Explorat. Newsletter, 1: 12-23. DOI: 10.1145/846183.846188.

Sumathi, C.P., Padmaja, R. Valli, 2010. Automatic Recommendation of Web Pages in Web Usage Mining,.International Journal on Computer Science and Engineering, (2): 9.

Xu, G., Y. Zhang and X. Zhou, 2005. Towards User Profiling for Web Recommendation. Proceeding of the 18th Australian Joint Conference on Artificial Intelligence (AI'2005), LNAI 3809, pp: 405-414, Australia.

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

Zhang Y and G. Xu, 2007. On Web Communities Mining and Analysis. 3rd international conference on Semantic, Knowledge and Grid (SKG2007), pp: 20-25, Oct 29-31, Xi'an, China, 2007.