2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Web Page Prediction by Clustering and Integrated Distance Measure

Poornalatha G[1], Prakash S Raghavendra[2]
Information Technology Department
National Institute of Technology Karnataka (NITK), Surathkal,
Mangalore, India
poornalathag@yahoo.com[1], srp@nitk.ac.in[2]

***Abstract*—The tremendous progress of the internet and the World Wide Web in the recent era has emphasized the requirement for reducing the latency at the client or the user end. In general, caching and prefetching techniques are used to reduce the delay experienced by the user while waiting to get the web page from the remote web server. The present paper attempts to solve the problem of predicting the next page to be accessed by the user based on the mining of web server logs that maintains the information of users who access the web site. The prediction of next page to be visited by the user may be pre fetched by the browser which in turn reduces the latency for user. Thus analyzing user's past behavior to predict the future web pages to be navigated by the user is of great importance. The proposed model yields good prediction accuracy compared to the existing methods like Markov model, association rule, ANN etc.**

Keywords-sequence alignment; user session; clustering;

I. Introduction

Web page prediction is the problem of forecasting the next page that might be visited by the user from the current active page or most recently visited previous pages. Various applications like web page recommendation, web site restructure, web caching and pre fetching, determining most appropriate place for advertisements, search engines etc. would benefit from the good prediction model. Consequently the web page prediction has gained more importance in recent years among research community. This paper proposes a prediction model that could be employed for above mentioned applications in general and web page prefetching in specific.

There are many architectures and related algorithms for developing web page predictor. Most of the researchers emphasize on the Markov model. Markov model is a mathematical tool for statistical modeling. The basic concept of the Markov model is to predict the next action, depending on the result of previous actions. Researchers have adopted this technique successfully in literature for training and testing the user actions and thus predicting their behavior in future. Deshpande et al. [1] discussed different techniques for selecting parts of different order Markov models to get a model with high predictive accuracy and less state complexity. The main idea was to eliminate some of the states of different order Markov models based on frequency, confidence and error. They predicted last page of the test session for evaluation purpose. Kim et al. [2] proposed a hybrid model by using Markov model, sequential association rule, association rule and a default model to improve the performance specifically the recall. But it did not improve the prediction accuracy. Khalil et al. [3] tried to improve the web page prediction accuracy by integrating clustering, association rules and Markov models. Dutta et al. [4] integrated first order Markov model with web page rank based on the link structure to get better prediction accuracy. Awad et al. combined Markov model with artificial neural network [5] and Support Vector Machines [6]. The final prediction was based on the Dempster's Rule. LRS was used to reduce the model complexity. Frequency matrix was used to represent first order Markov model.

J. Pitkow et al. [7] made an effort to reduce the model size compare to Markov models. But the static LRS model used by them may not be suitable for real time prediction model. Jalali et al. [8] [9] proposed LCS based algorithm for predicting user's future requests. They used graph partitioning algorithm for clustering and used LCS for classification. A new session is classified into any one of the clusters and prediction list is generated based on the navigation patterns of corresponding cluster. Gunduz et al. [10] proposed a prediction model by considering order information of pages and time spent on them in a session. User sessions are clustered by using graph partitioning algorithm and each cluster is represented by click stream tree. Mukhopadhyay et al. [11] proposed a clustering method to group related web pages based on access patterns. They used page ranking to build the prediction model in the initial stages. However the average hit percentage was around 50% for prediction window size of three and 35% for the prediction window with size two. Anitha [12] clustered the web log based on pair-wise nearest neighbor method, determined the next page accessed by sequential pattern mining. The sequence is not taken into consideration for clustering and Markov model is used for sequential mining. Lu et al. [13] generated significant usage patterns from the abstracted web session clusters by using Needleman-Wunsch global alignment algorithm. First order Markov model was build for each cluster. The accuracy of lower order Markov models are less and higher order Markov models consume space due to many states. Therefore, to improve the accuracy, researchers tried to integrate the Markov model with other data mining techniques like clustering, association rule etc.

978-0-7695-4799-2/12 $26.00 © 2012 IEEE
DOI 10.1109/ASONAM.2012.231

IEEE computer society

The objective of this paper is to propose a prediction model based on web user sessions clustering. Maintaining the information of order in which a user accesses web page of a web site is essential for predicting the next page to be accessed by the user. Hence, integrated distance measure based on the sequence alignment technique is used to compute the similarities between any two sessions. For this, SAM [15] and SABDM [16] methods are integrated.

The rest of the paper is organized as follows. The section II explains the proposed prediction model. The details of the experimental analysis are discussed in section III. The results obtained by the proposed model are given in section IV followed by references at the end.

II. Proposed Prediction Model

The proposed prediction model works in two phases namely offline and online phase as shown in Fig 1. The offline phase is carried out at the server whereas the online phase includes both server and the client. The various components depicted in this figure are discussed below.

Web log and Sessions-The offline phase considers server logs first. The log contains an entry for each of the access to the server by client. Each entry has information like client ip address, date and time at which the server is accessed, the HTTP method such as get, the requested URL, the response code, number of bytes transferred from server to client etc. Each entry is parsed to extract ip address, date and time, HTTP method and the URL requested. User sessions are created based on ip address, date and time. A user session consists of a set of contiguous sequence of web pages accessed by a user within certain amount of time (e.g. 30 minutes). Further the sessions are filtered to remove image files, robot navigations etc. Unique requests are identified from these filtered sessions and unique id is given to each of them. For example if there are 10 unique pages they are identified uniquely as P_1, P_2, P_3,…,P_{10}.

Clusters- User sessions are divided into training and testing sets. 60% of sessions are considered for training and the remaining 40% sessions are taken as the test data set. Clusters of training sessions are formed by using the modified k-means algorithm [14]. Initially cluster centers are selected in random as in standard k-means algorithm and the center may change in each of the iteration till the convergence. Once clusters are formed, unique identification is given to each cluster. For example if four clusters are formed they are uniquely identified as C_0, C_1, C_2 and C_3 respectively. The distance measure used to compute the similarities between two user sessions is discussed in section III B.

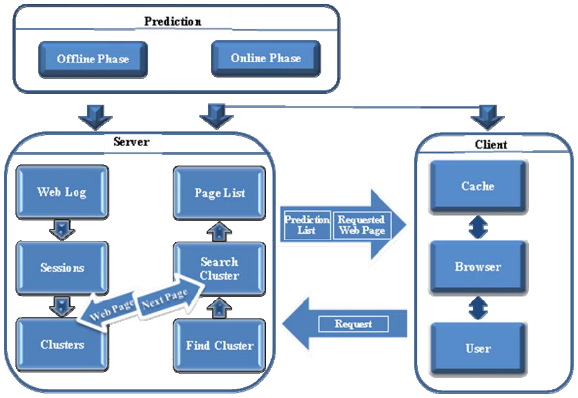

Figure 1. Prediction model

Client- Whenever user requests for a web page, the browser first looks into the cache for the presence of that web page. If found, the page is immediately retrieved and displayed to the user otherwise request is forwarded to the actual web server. The request consists of a set of pages accessed by the user. The server responds by sending the requested page followed by a prediction list. When the browser is idle, it tries to download the web pages listed in prediction list which are not presently available in the cache. When the user requests for the next page, the user immediately gets the page from local cache and the user gets response immediately. Thus by prefetching the web pages, the user latency could be minimized.

Find Cluster- This is the first stage of online phase. Here the nearest cluster to the request is determined by comparing it with all the existing cluster centers by using the integrated distance measure.

Search Cluster- This component initiates the search process in the clusters formed during the offline phase. The requested web page is searched based on the cluster obtained from the Find Cluster component. It retrieves the next page and maintains the count for each of the next page depending upon the number of sessions in which the next page is present followed by the current requested page in the cluster.

Page List- This component maintains list of pages received by the Search Clusters component. Further, the list can be pruned based on certain criteria. For instance, the list may be pruned based on frequency. That means top n web pages are selected based on highest frequency and prediction list is sent to client. For example, top 5 pages could be retained from page list. Thus after pruning the page list a prediction list is formed and sent to the client along with the response to the requested page.

Thus the proposed model predicts the next page possibly to be viewed by the user so that the pages could be pre fetched. This results in reducing the delay at the user end. Otherwise the user has to wait till the actual page is brought to him from the remote web server. The next section discusses about the experimental analysis done for the proposed prediction model.

III. Experimental Analysis

A. Data Set

The server log files are accessed only by the administrator of the web site and are not accessible to the common internet users. Therefore, it is difficult to get recent updated log files. Therefore one of the standard data sets used in the literature is taken for experimental purpose.

The data set considered is NASA log taken from http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html which consists of entries of July 1995.The data from the raw web log is parsed to extract relevant fields such as user ip address, date and time, requested URL, HTTP response code etc. Sessions are constructed after filtering unwanted entries like image files, response code with error, robot navigations etc. Further sessions with less than three pages are filtered assuming that they may not play significant role in prediction. Unique 860 requests are identified form these sessions and unique identities are assigned to them.

B. Integrated Distance Measure

This section briefly explains the SAM [15] and SABDM [16] distance methods. Also, the need for integrated approach is discussed. An example is given to understand the integrated distance measure approach to find the distance between any two user sessions.

The SAM [15] method partitions the navigation patterns based on the order of web pages requested by the user. Here identical pages of a pair of session are reordered to make them align with each other and unique pages are either inserted or deleted so as to make both sessions exactly similar to each other. The number of insertion operations, deletion operations and reorder operations are used to compute the distance between pair of sessions. The reorder operation alters the sequence in which the web pages are accessed in a session. The equation used in SAM method to find the distance between two sessions is given by (1).

$$d_{SAM}(S_1,S_2) = \min [(w_d D + w_i I) + \eta R] \qquad (1)$$

Where:

- d_{SAM} is the distance between two sessions S1 and S2, based on SAM
- w_d is the weight value for the deletion operations, a positive constant not equal to zero, determined by the researcher
- w_i is the weight value for the insertion operations, a positive constant not equal to zero, determined by the researcher
- D is the number of deletion operations
- I is the number of insertion operations
- R is the number of reordering operations
- η is the reordering weight, a positive constant not equal to zero, determined by the researcher

The equation (1) indicates that, the distance between two sessions consists of the costs for deletion and insertion of unique elements and the costs for reordering common elements. Since, deletion and insertion are single operations, the weights for w_d and w_i may be assumed as 1 and $\eta = w_d + w_i$. If the number of common pages viewed in both sessions is more but the order of page views is different, considerable amount of computation work is needed to find the number of common elements and reorder them and the reorder operation alters the original session. Also, SAM does not differentiate between the sessions that are completely different and sessions that have some similarities between them.

The Sequence alignment based distance method (SABDM) is described in [16]. Equation (2) is used to find the distance between two sessions.

$$d_{SABDM}(S_1,S_2)=[\max(m,n)-\text{similarityCount}]/\max(m,n) \qquad (2)$$

Where, m and n are length of two user sessions S_1 and S_2 respectively. similarityCount is number of alignments found by employing the local alignment algorithm [17]. The distance value lies in the range of 0 and 1. The value 1 indicates that the two sessions are completely different, where as 0 indicates that the sessions are exactly similar. Thus SABDM differentiates between sessions that are completely different. But, the direct alignment that exists between two sessions and the alignment obtained by inserting gaps is not differentiated in SABDM.

Therefore the integrated distance measure integrates both SAM and SABDM to get a better distance measure. As a result, the integrated approach finds the number of unique elements as in SAM and the number of alignments without changing the order of pages as in SABDM. Equation (3) gives the distance measure between two sessions S_1 and S_2. This distance measure essentially captures number of alignments between pair of original sessions, number of alignments obtained after local alignment and unique elements between them.

$$d_{INTEGRATED}(S_1,S_2)=[NUP+[2 \text{ x } (|NAP-NDA|)]]/(|S_1|+|S_2|) \quad (3)$$

Where:

- $d_{INTEGRATED}(S_1,S_2)$ is the distance between user sessions S_1 and S_2
- NUP (Number of Unaligned Pages) refers to the unique web pages of a pair of sessions that cannot be aligned by inserting gaps. For example, if S_1= (A, B, C, D) and S_2= (A, B, C, E) then D and E are unique and hence NUP is two; NUP is same as $(w_d D+w_i I)$ of SAM;
- NAP (Number of Aligned Pages) is the number of alignments found by employing local alignment method used in SABDM. NAP refers to the pages that could be aligned by inserting gaps. For example, if S_1=(A,B,C,F,G,H) and S_2=(A,F,G,J,U) then, by inserting two gaps after A in S2, F and G could be aligned to F and G of S1. Thus NAP is two.
- NDA (Number of Direct Alignments) refers to the pages that are aligned in the original pair of sessions. For example, if S_1=(A,B,C,F,G,H) and S_2=(A,E,G,F,U) then, NDA is two. Here, pages A and F are aligned between S1 and S2. The page A is present in the first position where as the page F is at the fourth position of both the sessions.
- $|S_1|$ is the length of session S_1
- $|S_2|$ is the length of session S_2
- |NAP-NDA| gives the actual number of pages that could be aligned by inserting gaps. This is multiplied by a constant 2 because two operations are performed while inserting a gap. i.e., inserting a gap is one operation and moving the existing page/s where gap is inserted is another operation.

To demonstrate the integrated distance measure, consider the example given below. Let S_1 and S_2 be two user sessions that represent the navigation path followed:

$S_1 = (P_1,P_2,P_5,P_6,P_3,P_1,P_2,P_5,P_4,P_8,P_7,P_2)$

$S_2 = (P_1,P_2,P_7,P_7,P_5,P_4,P_1,P_2,P_6,P_5,P_8,P_7,P_3)$

The alignments found between S_1 and S_2 are:

$$\begin{array}{ccccccccccccccc} P_1 & P_2 & - & - & P_5 & P_6 & P_3 & P_1 & P_2 & - & P_5 & P_4 & P_8 & P_7 & P_2 \\ | & | & & & | & & & | & | & & | & & | & | & \\ P_1 & P_2 & P_7 & P_7 & P_5 & P_4 & - & P_1 & P_2 & P_6 & P_5 & - & P_8 & P_7 & P_3 \end{array}$$

Here '-'indicates a gap and '|' indicates alignment. Since number of alignment obtained by inserting gaps are 6, NAP is 6. But, 1 and 2 are directly aligned and therefore NDA takes value 2. The unique pages left are 7, 7 and 2 according to SAM and thus NUP is 3. Length of sessions S_1 and S_2 are 12 and 13 respectively. Therefore the integrated distance, $d_{INTEGRATED}$ (S_1, S_2) is 0.44 by (3). Thus the integrated distance measure finds the distance between any two sessions that may be of uneven lengths by using the sequence information and without modifying the sequence.

C. Prediction

This section discusses the prediction done using the proposed model in detail. The proposed prediction model consists of two phases as depicted in Fig. 1 namely offline and online. In the first phase, sessions are created after preprocessing the entries of the server log file. The 60% of sessions called as train set are clustered by integrated distance method. In general a cluster contains some unique pages but there is a possibility that some pages like front page may be present in more than one cluster. Also, clusters are formed based on the sequence in which pages are accessed by the user and not based on the kind of pages accessed. For example, the page P_i may be accessed after P_j by some users and before P_j by some other users. So, the sessions containing P_iP_j and P_jP_i may be present in different clusters.

The online phase starts when the user requests a web page. The browser should first check the local cache for the required page. If the page is not available in the cache then the request along with the previous pages viewed by the user is forwarded to the remote web server. The server finds the cluster to which the request is nearest by applying the integrated distance as a similarity measure. The nearest cluster is explored for the presence of the last page of the request to find the next page accessed by various sessions and maintain a count for each of the next page. Thus given a request, the server prepares a unique list of URLs called Prediction List and sends to the client along with the response. The browser pre fetches the URLs in the prediction list during its idle time. The pages already present in the cache need not be pre fetched. Therefore the browser first checks the local cache for the presence of URLs of the prediction list and fetches those web pages that are not currently in cache. As a consequence, the delay at the user is reduced considerably because the page to be accessed by the user is now available locally if the prediction is correct.

IV. Results

Couple of experiments are conducted to evaluate the proposed prediction model by using the data set described in section III A. 5K, 10K and 15K number of sessions are considered for the experiment. To evaluate the accuracy of the predictions obtained by the proposed model, we use some definitions as discussed below:

- Hit – The predicted next page of a test session is present in the prediction list. Suppose (Pn1,Pn2,Pn3,Pn4) is a test session and (Pn1,Pn2,Pn3) is considered to find the nearest cluster 'C'. The cluster 'C' is searched for the presence of the page Pn3 and a prediction list is prepeared that contains pages that are viewed immediately after viewing Pn3 in 'C'. A hit results if Pn4 is present in the prediction list.
- Miss – The predicted next page of a test session is not present in the prediction list. Suppose (Pn1,Pn2,Pn3,Pn4) is a test session and (Pn1,Pn2,Pn3) is considered to find the nearest cluster 'C'. The cluster 'C' is searched for the presence of the page Pn3 and a prediction list is prepeared that contains pages that are viewed immediately after viewing Pn3 in 'C'. A miss occurs if Pn4 is not present in the prediction list.
- Match – The page or the path of a test session is found in a given cluster or the predicton list is not empty. Suppose (Pn1,Pn2,Pn3,Pn4) is a test session and (Pn1,Pn2,Pn3) is considered to find the nearest cluster 'C'. The cluster 'C' is searched for the presence of the page Pn3. If Pn3 is present in 'C' it is said to be a match.
- Mismatch – The page or the path of a test session is not found in a given cluster or prediction list is emty. Suppose (Pn1,Pn2,Pn3,Pn4) is a test session and (Pn1,Pn2,Pn3) is considered to find the nearest cluster 'C'. The cluster 'C' is searched for the presence of the page Pn3. If Pn3 is not present in 'C' it is said to be a mismatch.
- Accuracy – The ratio of number of correct predictions or hits to the number of matches. The number of mismatches are not considered for accuracy since the proposed model cannot predict for mismatches .

In the first experiment, the last page is predicted for the purpose of evaluating the proposed prediction model. Each session from the test data is considered one at a time. The last page of the test session is removed and the remaining pages were considered to find the nearest cluster center. For example, if a session consists of n number of pages $(P_1,P_2,\ldots,P_n)$, the last page P_n is removed. The remianing pages $(P_1,P_2,..,P_{n-1})$ were considered to find the nearest cluster. The nearest cluster obtained was searched for the presence of the page P_{n-1}. A prediction list is prepared that contains the next page visited followed by P_{n-1} in the cluster. The presence of actual last page of the session (P_n) is checked in the prediction list. A hit results if the prediction list contains P_n and the number of hits is incremented. Note

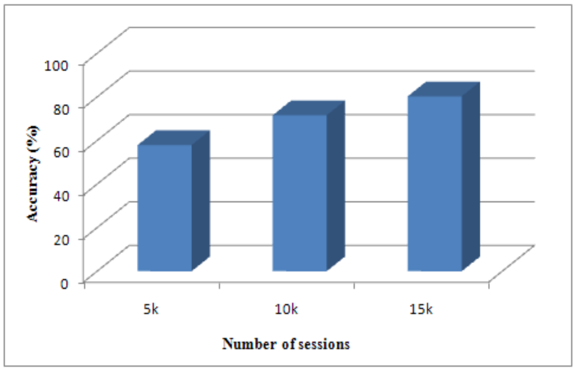

Figure 2. Prediction Accuracy

that, the pages from P_1 to P_{n-1} are considered to find the nearest cluster. But, once the nearest cluster is found, only the last page i.e., P_{n-1} is used to predict the next page. Fig. 2 depicts the accuracy for the NASA data set.

Table 1 shows the number of hits, misses, matches, mismatches and average session length by considering different number of sessions. It can be observed that, the overall number of matches is more than the mismatches. The numbers of hits are more than the misses as the size of training data increases. The better % of matches indicates the goodness of the prediction model as well as the integrated distance measure which is used to find the distance between the test sessions to the nearest cluster.

The second experiment is conducted to limit the list size and analyze the accuracy of predictions. The page list is sorted in descending order based on the frequency and the top n pages from this sorted page list are selected to form the prediction list. The test data set is evaluated by varying the n values from 1 to 10. Fig. 3 depicts the accuracy for these top n values. The graph illustrates the improvement in accuracy as the number of pages in the prediction list increases. Instead of providing a prediction list with all the pages, it would be better to have 'n' number of pages in the prediction list as this list is used for prefetching pages from the client browser. Depending on the bandwidth available or the server load, suitable value for 'n' could be assumed.

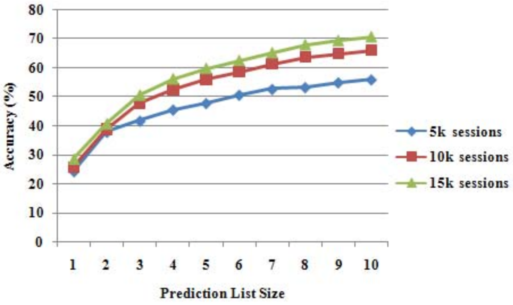

Figure 3. Prediction accuracy for various list sizes

TABLE I. RESULTS OF PREDICTION

Number of sessions	5K	10K	15K
Hit (%)	57.91	71.58	80.29
Miss (%)	42.09	28.42	19.71
Match(%)	87.90	97.20	98.17
Mismatch(%)	12.10	2.80	1.83
Average session length	8.39	8.94	8.62

The authors in [5] compare prediction results of various methods like Markov, ANN, ARM, All-kth-Markov, All-kth-Dempster's rule etc. by considering ranks for web pages based on highest confidence. The probability of hit by match is not more than 54% for ranks 1-8 for various techniques. The figure 3 clearly shows that, the accuracy achieved by the proposed model is much better than the accuracy obtained by All-kth- Dempster's rule in [5]. The proposed model ranks only the predicted web pages based on highest frequency whereas, frequency matrix is constructed in [5]. The entries of this matrix represent frequency of two consecutive pages for each of the web page. As the number of pages in a web site will be more, the frequency matrix consumes more space. The accuracy of prediction increases as the numbers of hops are increased. But, the proposed method achieves better accuracy with only one hop. Hop is the number of pages considered to predict the next page in [5].

F.Khalil et al. [18] combined Markov model, association rules and clustering to predict the next page to be accessed by user. They used 4 different data sets to evaluate their model. However, the accuracy obtained was 45%, 55%, 65% and 35% respectively for four data sets used. J.Pitkow [7] tested their prediction model by comparing one-hop LRS with one-hop Markov model. The probability of hit by match was 25% and for k-th order Markov model was 31%. Y.Z.Guo et al. [19] ranked web pages based on access time, length and frequency. The prediction accuracy obtained was around 50% for top 3 predictions and less than 60% for top 5 predictions. Thus, the other methods in literature mentioned above do not accomplish good accuracy compare to the proposed prediction model.

V. CONCLUSIONS AND FUTURE WORK

A prediction model that yields good accuracy is proposed in this paper. Also, an integrated distance method is proposed to find the similarities between any two user sessions based on the sequence alignment. The results obtained are compared with few other results available in the literature to demonstrate the goodness of the prediction model proposed. Since accuracy is good, the proposed work could be useful for prefetching applications that reduce the user latency.

The future work may consider predicting next page based on sliding window. One more data set could be used to evaluate the proposed work along with the NASA data set so that the consistency of results could be analyzed with more number of data sets.

References

[1] M. Deshpande, G. Karypis, “Selective Markov Models for Predicting Web Page Accesses,” ACM transactions on Internet Technology, vol. 4, No.2, pp.163-184, May 2004

[2] D. Kim, l. lm, N. Adam, V. Atluri, M. Bieber, Y. Yesha, “A Clickstream-Based Collaborative Filtering Personalization Model: Towards A Better Performance,” Proceedings of the 6th annual international workshop on web information and data management, ACM, pp.88-95, 2004. DOI:10.1145/1031453.1031470.

[3] F. Khalil, J. Li, H.Wang, “Integrating Recommendation Models for Improved Web Page Prediction Accuracy,” Proceedings of the thirty-first Australian conference on Computer Science, vol.74, Jan 2008

[4] R. Dutta, A. Kundu, R. Dattagupta, D. Mukhopadhyay, “An Approach to Web Page Prediction Using Markov Model and Web Page Ranking,” Journal of Convergence Information Technology, vol.4, Dec 2009

[5] M. A. Awad, L. R. Khan, “Web Navigation Prediction Using Multiple Evidence Combination and Domain knowledge,” IEEE transactions systems, man and cybernatics-part A:systems and humans, vol.37, no.6, pp.1054-1062, Nov 2007

[6] M. A. Awad, L.R. Khan, “Predicting WWW surfing Using Multiple Evidence Combination ,” The VLDB Journal , Springer-Verlag , vol.17, pp.401-417, 2008

[7] J. Pitkow, P. Pirolli, “Mining Longest Repeating Subsequences to Predict World Wide Web Surfing,” Proceedings of the 2nd USENIX Symposium on Internet Technologies & Systems, pp.13-13, Oct 1999

[8] M. Jalali, N. Mustapha, Md. N. B. Sulaiman, A. Mamat, “A Web Usage Mining Approach Based on LCS Algorithm in Online Predicting Recommendation Systems,” 12th International Conference on Information Visualisation, IEEE., pp.302-307, 2008 DOI: 10.1109/IV.2008.40.

[9] M. Jalali, N. Mustapha, A. Mamat, Md. N. B. Sulaiman, **”** A new classification model for online predicting users’ future movements,” International symposium on Information Technology, , IEEE, 2008 DOI: 10.1109/ITSIM.2008.4631852

[10] S. Gunduz, M.T. Ozsu, “A Web Page Prediction Model Based on Click-Stream Tree Representation of User Behavior,” Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Aug 2003 doi:10.1145/956750.956815

[11] D. Mukhopadhyay, P. Mishra, D. Saha, Y. Kim,”A Dynamic Web Page Prediction Model Based on Access Patterns to Offer Better User Latency,” The 6th International Workshop MSPT 2006 Proceedings; Youngil Publication; Republic of Korea, pp. 59–64, November 20, 2006

[12] Anitha, “A New Web Usage Mining Approach for Next Page Access Prediction,” International Journal of Computer Applications, vol.8, no.11, October 2010

[13] L. Lu, M. Dunham, Y. Meng, “ Discovery of significant usage patterns from clusters of clickstream data,” *WebKDD '05*, ACM, pp. 139-142, 2005

[14] G. Poornalatha, P. S. Raghavendra, “ Web User Session Clustering Using Modified K-means Algorithm,” First International Conference on Advances in Computing and Communications (ACC – 2011), CCIS(191),Springer-Verlag, pp.243-252, 2011

[15] B. Hay, G. Wets, K. Vanhoof, “Mining Navigation Patterns Using a Sequence Alignment Method,” Journal of Knowledge and Information Systems, Springer-Verlag, pp. 150-163, 2004

[16] G. Poornalatha, P. S. Raghavendra, “Alignment Based Similarity Distance Measure for Better Web Sessions Clustering,” Journal of Procedia CS, vol.5, pp. 450-457, 2011 doi:10.1016/j.procs.2011.07.058

[17] S. Temple, W. Michael, “ Identification of Common Molecular Subsequences,” Journal of Molecular Biology, vol.147, pp. 195-197, 1981

[18] F.Khalil, J. Li, H. Wang, “An Integrated Model for Next Page Access Prediction,” Inderscience Enterprises Ltd., 2009

[19] Y. Z. Guo, K. Ramamohanarao, L. A. F. Park, “Personalized PageRank for Web Page Prediction Based on Access Time-Length and Frequency, ” IEEE/WIC/ACM International conference on Web Intelligence, pp.687-690, 2007