Web Page Access Prediction Using Hierarchical Clustering Based on Modified Levenshtein Distance and Higher Order Markov Model

Harish Kumar B T,
Assistant Professor, Dept Of CSE,
BIT, Bangalore, India

Dr. Vibha L,
Professor, Dept of CSE,
BNMIT, Bangalore, India

Dr. Venugopal K R,
Principal,
UVCE, Bangalore, India

***Abstract*--- Web Page access prediction is a challenging task in the current scenario, which draws the attention of many researchers. Predictions need to keep track of history data to analyze the usage behavior of the users. Web Usage behavior of a user can be analyzed using the web log file of a specific website. User behavior can be analyzed by observing the navigation patterns. This approach requires user session identification, clustering the sessions into similar clusters and developing a model for prediction using the current and earlier accesses. Most of the previous works in this field have used K-Means clustering technique with Euclidean distance for computation. The drawbacks of K-Means is that deciding on the number of clusters, choosing the initial random center are difficult and the order of page visits are not considered. The proposed research work uses hierarchical clustering technique with modified Levenshtein distance, Page Rank using access time length, frequency and higher order Markov model for prediction. Experimental results prove that the proposed approach for prediction gives better accuracy over the existing techniques.**

Keywords--- Access Time; Frequency; Hierarchical Cluster; Levenshtein Distance; Markov Model; Prediction

I. INTRODUCTION

Web is an important medium of communication and information exchange. The amount of information in the web is increasing every day and the demand for the information is also proportionally increasing. Websites provide web log files which are automatically created and maintained by the web servers. Every access to the website, including each view of the HTML document, image or other object is logged. The raw web log file format is one line of text for each access to the website. This contains information about who visited the site, where they came from, what they were doing on the website and how long they have accessed the page. Web log file are of two types:

- Common log format
- Extended log format or Combined log format

A. Common Log Format

Old web servers used this format. The common log format contains the fields such as Client IP Address, User id, Access Date and Time, HTTP Request Method, Path of the Resource on the web server, Protocol used for Transmission, Status code for the Transmission and Number of bytes transmitted.

Eg: 202.244.227.66 - - [01/Jul/1995:05:25:36 -0400] "GET http://www.enggresources.com/shuttle/missions/missions.html HTTP/1.0" 200 12054.

B. Extended Log Format

The recent web servers like Apache supports extended log format by inserting additional fields like referrer and user agent.

Referrer: Contains the link from where the user has arrived at the required website.

User Agent: Contains the browser used by the user to access the web site.

Eg: 202.244.227.66 - - [01/Jul/1995:05:25:36 -0400] "GET http://www.enggresources.com/shuttle/missions/missions.html HTTP/1.0" 200 12054 htt://www.google.com Mozilla.

C. Motivation

Thirst for knowledge and increasing demand for accessing information slows down the performance of satisfying the request made by the users, hence page prediction technique allows prefetching the pages before the request is made by the user and the performance of fetching the page can be maintained. Page prediction will predict where the user might visit next and preload that page ahead of time. This makes the navigation faster. Page prediction involves analysing and understanding the usage pattern for producing the useful information.

D. Contribution:

In the proposed work, web page prediction technique is proposed which involves web log data cleaning and conversion, user and session identification, clustering the user and sessions using hierarchical clustering and higher order Markov model is used for prediction.

The remaining sections of the paper are organized as follows. Section II gives the overview of the related work in the specified research area. In section III problem statement, aims and objectives are discussed, section IV presents the general architecture of the proposed work. Proposed methodology and working example is presented in section V. Section VI contains the algorithms for the proposed

978-1-5090-0931-2/16/$31.00 ©2016 IEEE

methodology. Section VII contains experimental setup and performance analysis. Section VIII contains conclusions.

II. RELATED WORK

Lot of research work is happening in the field of web page prediction. A brief survey of the related work in this research direction is presented below.

Phyu Thwe in [1] has used the K-means clustering technique to cluster the data sets after pre-processing, second order Markov model, popularity and similarity based page rank algorithm to make prediction. Y.Z Guo et al., in [2] has proposed personalized page rank for web page prediction based on access time length and frequency of access using higher order Morkov model.

In [3] Smrithi Pandya et al., has presented the review paper on web page prediction and has discussed several data mining approaches for web page prediction. Clustering the data set is essential for processing them in an effective way in [4] Tingzhong Wang has proposed a methodology for web log mining based on improved K-Means clustering which reduces the outliers and improves the clustering results.

Prediction helps in prefetching the web pages and eliminates the latency problems. Sonia Setial et al., in their survey paper [5] have dealt with several techniques for web prefetching namely, Markov model based technique, prediction by partial match based approach, double dependency graph based approach, sequential mining approach, clustering method and content based technique.

Analysing the web user's navigational behaviour is important task for predicting the next web page. Bindu Madhuri et al., in [6] has proposed a method to analyse the navigational behaviour of users using GRPA (Grey Relational Pattern Analysis) and VLMC (Variable Length Markov Chains).

Sweah Liang Yong in [7] has discussed about Web page prediction using the machine learning approaches by creating Graph Neural Network (GNN) which can successfully learn many other web page ranking methods like TrustRank and HITS. Markov models have been widely used to represent and analyse user's navigational behaviour (usage data) in web graph using the transitional probabilities between web pages, as recorded in the web logs. The recorded user's navigation is used to extract popular web paths and predict current user's next steps. In [8] Sonal Vishwakarma has analysed and studied k^{th} order Morkov model with webpage keywords as a feature to give more accurate results in web page prediction.

S Prince Mary et al., in [9] has proposed the details of web log pre-processing steps and identifying the usage patterns. In [10] a new method for next page prediction using PNN (Pair Wise Nearest Neighbour) and Sequential Pattern mining technique has been proposed.

III. PROBLEM STATEMENT

PageRank algorithms are highly biased towards the new web pages as they might not have many in-links at the initial stages. Web Page prediction systems are based only on the usage and ignore the Page-Ranking. To reduce the web page access latency, predicting the current user's next move is essential. An integrated approach to predict the next page based on the Page-rank using access time and frequency of access (TFPR), Clustering and Higher Oder Markov Model is desired.

Objectives:

- Creating the web log data attributes grouped by user and time.
- Form clusters by finding the similarity between each pattern.
- Finding the Transition Probability Matrix (TPM), Access Time with TPM (ATPM) and Frequency of access with ATPM (FATPM).

IV. ARCHITECTURE AND MODELLING

Web log file from the NASA website is used as an input to the proposed work. Architecture and Modelling of the current system is depicted in Fig. 1 which consists of three modules Pre-Processing, Hierarchical Clustering and Markov Model for prediction.

V. PROPOSED METHODOLOGY

Web page prediction means anticipating the next page that will be accessed by the user when browsing a website. User's previous website navigation pattern is very crucial for extracting useful information necessary for predicting the future click or access by the current user. An example, if 95% of the users' access page X after accessing pages U, V, W then there is a likely 95/5 probability that the current user accessing pages U, V, W will also access page X next. The proposed methodology aims to improve the prediction accuracy by combining the Hierarchical Clustering, TFPR and Markov Modelling.

A. Hierarchical Clustering

Hierarchical Clustering is a method in data mining which allows clustering analysis. Hierarchical clustering can be achieved in two ways.

- Agglomerative Hierarchical Clustering
- Divisive Hierarchical Clustering

The proposed methodology uses the Agglomerative Hierarchical Clustering. To cluster the sessions, measuring the similarity between the sessions is necessary. This is

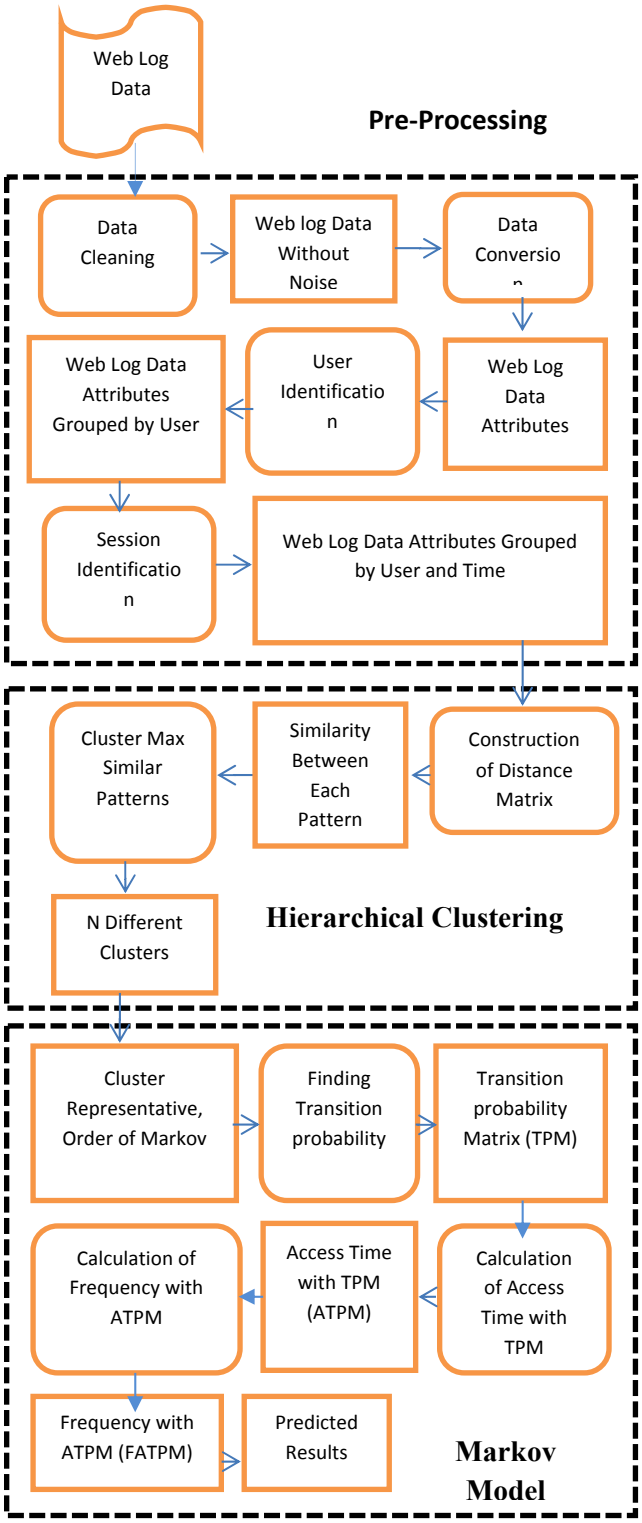

Fig. 1: Architecture and Modelling

achieved by using an appropriate distance metric. Modified Levenshtein distance is used as the distance metric which is an improved form of Edit distance (Levenshtein Distance). Levenshtein distance has a severe drawback when applied to the web sessions. It does not consider the page sequence into account. Page sequence is very essential in clustering and prediction.

Modified Levenshtein distance technique can be used for checking the string similarity and it also takes into account the page visit sequence. Levenshtein distance is modified to make best technique for forming the hierarchical clustering of the web sessions. Modified Levenshtein distance equation is as shown in (1).

$$M(i, 0) = 0,\ 0 \le i \ge L1 \text{ and } M(0, j) = 0,\ 0 \le j \ge L2$$

Where,
$M(i, j)$ is maximum similarity score matrix.
$L1$ and $L2$ are lengths of session S_i and S_j respectively.

$$M(i, j) = \text{Max}\{0, M(i-1, j-1) + 2 \text{ if } S_i = S_j \text{ else } -1, M(i-1, j) + W_d, M(i, j-1) + W_i\}$$
$$\text{for } 1 \le i \ge L1,\ 1 \le j \ge L2 \qquad (1)$$

where,
$w_i = -1$ (insertion penalty) and $w_d = -1$ (deletion penalty)

Maximum similarity score matrix construction and working of Modified Levenshtein Distance similarity score is as shown below. Consider the example session patterns as show in Table 1. There are seven sessions.

TABLE 1: Session Patterns

Session ID	Access Pattern
S1	P1,P2,P3,P4,P5
S2	P4,P5
S3	P1,P2,P5
S4	P6,P7
S5	P1,P2,P3,P4,P5
S6	P5,P6,P7
S7	P5,P6

B. Maximum Similarity Computation

Example 1 shows the steps to compute the maximum similarity between the sessions S1 and S2 and the same is represented as maximum similarity matrix as show in Table II. The maximum entry in the matrix shown in Table II is at M (5, 5) = 4. The maximum similarity between S1 and S2 is 57% computed using (2).

Example 1 : Session (S1, S2):
$L1 = 5$ (Length of S1) and $L2 = 2$ (Length of S2)
For $1 \le i \ge L1,\ 1 \le j \ge L2$

When $i = 1$ and $j = 1$
$M(1,1)=\text{Max}\{0,M(0,0)-1,M(0,1)-1,M(1,0)-1\}$
$M(1,1)=\text{Max}\{0,0-1,0-1,0-1\}$
$M(1,1)=0$

TABLE II: Similarity Matrix of S1 and S2

	-	P1	P2	P3	P4	P5
-	0	0	0	0	0	0
P4	0	0	0	0	2	1
P5	0	0	0	0	1	4

$$\text{Max Similarity} = \frac{Max}{L1+L2} * 100 \quad (2)$$

Where,
Max = Maximum Entry in Similarity Matrix.
L1 = Length of S1 and L2 = Length of S2

Table III contains the maximum similarity score between all pairs of raw sessions show in Table I. Fig. 2 shows the dendogram representation of the hierarchical cluster with sessions on x axis and dissimilarity on y axis.

TABLE III: Maximum Similarity score

	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**
S1	-	0.57	0.50	0.00	1.00	0.00	0.28
S2	0.57	-	0.40	0.00	0.57	0.40	0.50
S3	0.50	0.40	-	0.00	0.50	0.33	0.40
S4	0.00	0.00	0.00	-	0.00	0.80	0.50
S5	1.00	0.57	0.50	0.00	-	0.25	0.28
S6	0.00	0.40	0.33	0.80	0.25	-	0.80
S7	0.28	0.50	0.40	0.50	0.28	0.80	-

C. *Higher Order Markov Model*

Markov Model is a stochastic process with the Markov property which specifies that the probability of the next state depends only on the probability of the current state. Markov model which gives the probability of the next state using the current state is called as First Order Markov model. First order Markov model is memory-less as it does not remember the previous states except the current state. Predicting the next state only using the current state is very difficult. Generalized higher order Markov model for 2nd, 3rd and nth order is as shown below.

$X_{n+1}=P(X_{n+1}=j|X_n=i_n,X_{n-1}=i_{n-1})$ ------ 2nd Order
$X_{n+1}=P(X_{n+1}=j|X_n=i_n,X_{n-1}=i_{n-1},X_{n-2}=i_{n-2})$ ------ 3rd Order
$X_{n+1}=P(X_{n+1}=j|X_n=i_n,X_{n-1}=i_{n-1},X_{n-2}=i_{n-2},\ldots,X_0=i_0)$---nth Order

To predict the future, a little bit of past memory is required. Transition Probability Matrix (TPM) is constructed by taking all the page categories into account. Transition Probability Matrix is computed as depicted in (3).

$$P[i,j] = n[i,j] \,/\, \sum_{j=0}^{n} n[i,j] \quad (3)$$

Here, n[i,j] is the number of times there is transition from

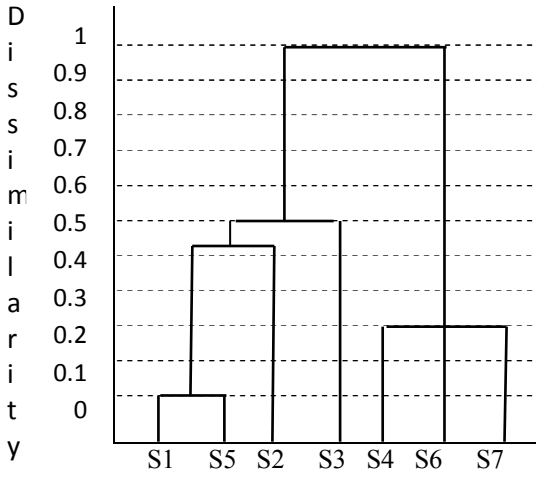

Fig. 2: Dendogram Representation

$page_i$ to $page_j$. P[i,j] is the probability of transition from $page_i$ to $page_j$.

Where,

$$P[i,j] \geq 0 \text{ and } \sum_{j=0}^{\infty} P[i,j] = 1 \;\forall\; i,j$$

For the example considered there are 7 page categories, hence, a Transition Probability Matrix of 7X7 has to be constructed as shown in Table IV for the cluster which contains sessions S1, S5, S2 and S3. Page sequence is P1, P2, P3, P4, P5, P1, P2, P3, P4, P5, P4, P5, P1, P2, P5. The inference from the matrix is that the probability of accessing P2 after P1 is 1.00, probability of accessing P3 after P2 is 0.66 and probability of accessing P5 after P2 is 0.33. Transition Probability matrix is multiplied with access frequency and access time length matrix.

TABLE IV: Transition Probability Matrix

	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**
P1	0.00	**1.00**	0.00	0.00	0.00	0.00	0.00
P2	0.00	0.00	**0.66**	0.00	**0.33**	0.00	0.00
P3	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00
P4	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00
P5	**0.66**	0.00	0.00	**0.33**	0.00	0.00	0.00
P6	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P7	0.00	0.00	0.00	0.00	0.00	0.00	0.00

VI. ALGORITHMS

The various algorithms proposed in this work are shown in this section. Data cleaning and conversion algorithm shown in Table V takes the raw weblog file as the

input and outputs the relevant records that are saved into the database.

TABLE V: Algorithm for web log cleaning and conversion

Algorithm : Data cleaning and conversion
Input : Weblog file
Output : Relevant records saved in DB (nasalog)
Method:
for each Record in weblog file
  Read fields
    If fields = {*.gif,*.jpg,*.css} OR {404,500 } then
      >> Remove Records
    Else
      >> Save Records in nasalog table
    End if
Until no more records

Algorithm extracts each fields form the web log and checks if the request is for *.gif, *.jpg, *.css or unsuccessful request and removes those records. All other records are stored in the relational database for future processing.

TABLE VI: Algorithm for Identify User and Session

Algorithm : User and Session identification
Input : Relevant records saved in DB nasalog
Output : Set of sessions
Method:

- For each record in db
- Repeat steps
  - Compare ip-address of first entry with ip-address of second entry.
  - If both are same identify both entries are from same user.
- Until last entry
- For each user,
  - Order records by time
  - identify <=30min entry from first entry of web page && minimum 5 page views in a session
- Until last entry

Algorithm for identifying the user and computing the sessions is as shown in Table VI. Relevant records saved in the nasalog database are given as input to this algorithm. Set of sessions are obtained as output for session time of maximum 30 minutes with minimum 5 page views per session.

TABLE VII: Algorithm for Hierarchical Clustering

Algorithm : Hierarchical Clustering
Input : Set of sessions stored in sessions DB
Output : k clusters
Method:

- Construct Modified Levenshtein distance matrix between every pair of session as shown in (1).
- Distance is number of similar pattern which maintain order of navigation
- Find the Maximum Element in the Matrix as shown in Table II.
- Compute the Maximum Similarity Score Using (2)
- Merge highest similar patterns together and their sessions as shown in Fig. 2
- Update the session list
- Merge until no more no more merging possible

Set of sessions obtained from the user and session identification algorithm are given as input to the Hierarchical clustering algorithm shown in Table VII to get *k* clusters. Modified Levenshtein distance shown in (1) is used to find the similarity between the sessions and sessions with higher similarity are merged hierarchically.

TABLE VIII: Markov Model Algorithm

Algorithm : Markov model
Input : Cluster representative
Output : Transition probability matrix
Method:

- Construct Transition Probability Matrix (TPM) of *nXn*, where *n* is no. of page category.
- Calculate probability for each state and update the TPM using (3)
- Multiply TPM with TPM of n-times to get n-order Markov model.
- Probability distribution will be used for prediction

Markov Model algorithm is shown in Table VIII. Cluster representative is the input and gives the Transition Probability Matrix which gives the probability of moving to the next page from the current and previous pages.

VII. EXPERIMENTAL SETUP AND PERFORMANCE ANALYSIS

The proposed work aims to improve the web page access prediction accuracy by using hierarchical clustering and higher order Markov model. To evaluate the performance of the proposed methodology NASA web access log file is considered. Web log data of 5 days from 01/July/1995 to 05/july/1995 was taken for training and one day log data (6/july/1995) for testing the accuracy of prediction. Raw data considered for training consisted of 3,79,582 entries. Cleaning operation was then performed on the raw data, which then reduced to 80,329 entries with 19,571 distinct users and 562 distinct pages. 4051 user sessions were obtained using the session timeout period as 30 minutes and a minimum of 5 page views in a session. The similarity score between every pair of sessions was computed using modified Levenshtein distance, hierarchical cluster formation was done with 524 clusters at 93 different levels. Probability matrix of the order 562 by 562 was constructed and probability of accesses was computed for various orders of Markov. Test data consisted of 1,00,960

records and after filtering reduced to 21,125 records with 445 distinct pages and 6,103 unique users. 1,141 sessions were obtained for 30 minutes sessions timeout and minimum 5 page views per session.

Performance of the proposed methodology was evaluated at different levels of hierarchy in a hierarchical cluster model and it has been observed that prediction accuracy improves with the decrease in hierarchy level (because more similar sessions are grouped at higher levels). Fig. 3 shows the prediction accuracy for different levels. This concludes that the similarity level of 65% to 75% gives the better accuracy.

Proposed methodology is also compared with three other existing methodologies such as Time and Frequency based Page Rank (TFPR), Usage based ranking (UBR) and Prediction using K-means clustering with Euclidian Distance and Markov Model (KEDM). Fig. 4 shows the prediction accuracy for existing techniques and the proposed technique. It has been observed that the proposed methodology using hierarchical clustering based on modified Levenshtein distance, higher order Markov model, Page frequency and Access time length yields an accuracy of 65% to 75% which is the best.

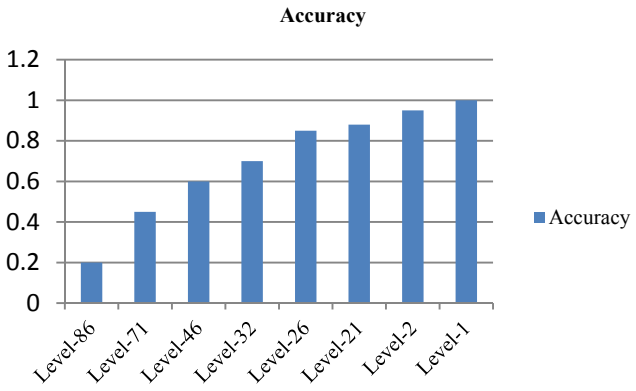

Fig. 3: Accuracy at Different Similarity Levels

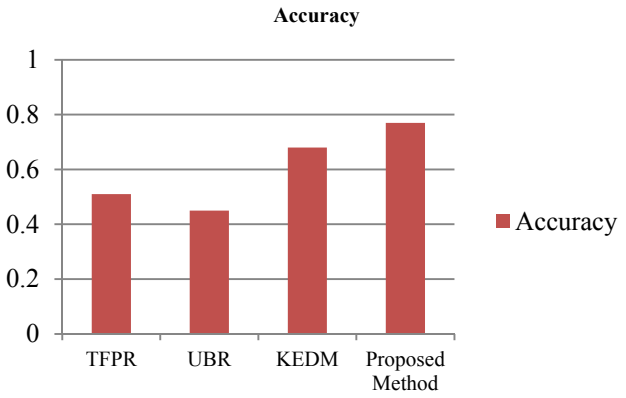

Fig. 4: Performance comparison with existing Techniques

VIII. CONCLUSION

In the proposed work a new methodology using hierarchical clustering based on modified Levenshtein distance and higher order Markov model is proposed to improve the web prediction accuracy. The proposed work can be used to prefetch the web pages before they are actually being requested by the user, this reduces the access latency.

REFERENCES

[1] Phyu Thwe, "Using Markov Model and Popularity and Similarity Based PageRank Algorithm for Web Page Access Prediction", *International Conference on Advances in Engineering and Technology (ICATE),*March 29th-30th, 2014, Singapore.

[2] Y.Z Guo et al., "Personalized PageRank for Web Page Prediction Based on Access Time Length and Frequecy", *In Web Intelligence, IEEE/WIC/ACM International Conference,* Page 687-690.

[3] Smriti Pandya et al., "Review Paper on Web Page Prediction Using Data Mining", *International Journal of Computer Engineering and Intelligent Systems,* Vol 6, No. 7, ISSN 2222-1719 (Paper), ISSN 2222-2863 (online)-2015.

[4] Tingzhong Wang, "The Development of Web Log Mining Based on Improved K-Means Clustering Analysis", *Advances in CSIE,* Vol 2, AISC 169, PP 613-618 Springer Verlag Berlin Heidelberg, 2012.

[5] Sonia Setia et al., "Survey of Recent Web Prefetching Techniques", *International Journal of Research in Computer and Communications Technology,* Vol 2, Issue 12, Dec-2013 ISSN: 2278-5841 (Online), ISSN: 2320-5156 (Print)

[6] Bindu Madhuri et al., "Analysis of User's Web Navigation Behaviour Using GRPA with Variable Length Markov Chains", *International Journal of Data Mining and Knowledge Management Process (IJDKP),* Vol 1, No. 2, March-2011.

[7] Sweah Liang Yong, "Ranking Web Pages Using Machine Learning Approaches", *In International Conference on Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM,* 2008.

[8] Sonal Vishwakarma, "Web User Prediction By: Integrating Markov Model with Different Features", *International Journal of Engineering and Science and Technology,* Vol 2, No. 4, Nov-2013, ISSN: 2319-5991.

[9] S. Prince Mary et al., "An Efficient Appraoch to Perform Pre-Processing", *International Journal of Computer Science and Engineering,* Vol 4, No. 5, Oct-Nov 2013, ISSN: 0976-5166

[10] A. Anitha, "A New Web Usage Mining Approach for Next Page Access Prediction", *International Journal of Computer Application,* Vol 8, No. 11, Oct-2010, ISSN: 0975-8887.