Prediction of Web User's Browsing Behavior using All Kth Markov model and CSB-mine

Neha V. Patil[#1], Dr. Hitendra D.Patil[*2]

[#1]*Master Student, Department of Computer Engineering, SSVPS'S B.S.Deore College of Engineering, India*
[*2]*Professor and Head, Department of Computer Engineering, SSVPS'S B.S.Deore College of Engineering, India*

Abstract- *Web Usage Mining (WUM) provides interesting rules and patterns of user web browsing behavior and prediction technique is very useful where the next set of web pages are predicted based on history of visited web pages. Nowadays use of the web and so the internet traffic has increased, so prediction of user web behavior plays important role in various applications like smart phones, recommendation systems and web personalization, etc. Various prediction models have been proposed based on Markov models, Association Rule Mining (ARM), etc. The Specific order of the Markov model cannot predict for a session that was not observed previously in the training set. ARM endures scalability issue, originates from generating item sets. Proposed system first preprocess raw web log, construct sessions for different users, then All Kth Markov model and using Conditional Sequence Base (CSB-mine), Sequential Access Patterns based model both are used individually to predict the next page that a user may visit.*

Keywords: *WUM, Web session, Web browsing behavior, web prediction, All Kth Markov, CSB-mine*

I. INTRODUCTION

The World Wide Web (WWW) is growing at high speed in terms of network traffic and the size of the Web sites. As a result, how to provide Web users with more exactly required information is becoming a crucial issue in Web applications, so it is very important to identify useful web data to capture the interests of users. In Web mining, data mining techniques are used toextract useful knowledge from web data, consisting web documents, web sites access logs, web page structure, links between documents, etc. Web data mining is classified in three categories as Web Content Mining, Web Structure Mining and Web Usage Mining. Extraction of required data, information and knowledge available in the web page is referred as Web Content Mining. Web structure mining focuseson analysis of the web site's hyperlink structure. Web usage mining process deals with the extraction of interesting usage patterns from web log data[2]. This web log data is the web user access logs when a user accesses the web server.Web log data collected on a Web server include IP addresses, domain name, page references, and access time of the users, browser information etc. The extracted data from Web usage mining can be useful in many web applications such as web caching, web page recommendation, search engines, web site restructuring and personalization [3]. Web usage mining consists of three tasks, namely data pre-processing, pattern discovery, and pattern analysis. In data preprocessing raw data is filtered to get the required data, then patterns, rules and statistics are figure out so that useful rules and patterns are determined in pattern analysis phase. These interesting patterns are very useful to predict web users browsing behavior.

In Web Prediction the knowledge of previously accessed web pages from user web surfing behavior which is maintained on the web server are very useful. For many businesses web is the most important medium for marketing and sales,so the accurate prediction of Web navigation patterns plays a crucial role. Often these predictions are based on complex temporal models of users' behavior learned from historical data. Significant patterns do exist in Web navigation data. Learning and predicting such patterns have immense commercial value as the Web evolves into a primary medium for marketing and sales for many businesses. Web-based businesses look for useful users' patterns to identify promising events, potential risks, and to undertake customer relations management, also to help them optimize their business processes and system operations.

A history of previously accessed web pages of different users is used to predict a future set of pages likely to be visited by a user. Such knowledge of user's previously accessed web page navigation within a period of time is called as a web session. It gives an exact accounting of who accessed the Web site, what pages were requested and in what order, and how long each page was viewed. All of the web page accesses that occur during a single visit to a Web site are referred as a user session. These web sessions play important role in web prediction and are an importantsource of data for training, and they made up of sequences of web pages that users have visited along with date, time. Improving the prediction process can be very useful to decrease network traffic by avoiding accessing unnecessary web pages. Basically web prediction is considered as a classification task in which next pages are predicted using the history of previously visited web pages [1].

In this paper, the work follows processes as Data Preprocessing which cleans web log data and identifies all users' web sessions since the information contained in a raw Web server log does not reliably represent a user session file. All Kth Markov model and CSB-mine to generate Sequential Access Patterns (SAP) are constructed and both the models are then individually used for prediction of next web page that the user may visit. The system is evaluated using web server log of NASA Kennedy Space Center [13].

The organization of this paper is as follows. In Section II, we present the work related to the different prediction models. Section III describes the analysis of the limitations of the existing systems. The working of the proposed system is discussed in section IV. Section V discusses the details of the experiment withthe results. Finally, in section VI we discuss conclusion.

II. RELATED WORK

There are different conventional prediction models to predict web browsing behavior of the user. This section describes these prediction methods. Prediction models are divided into two categories as point based and path based models. In point based prediction model, prediction is performed depending on currently observed user actions. In the path based prediction model,prediction is performed depending on the user's previous path data i.e. previous navigation patterns. As path based models are able to look far in the user's history of accessed web pages, they achieve more accuracy compare to point based models.

T. Joachims, D. Freitag, and T. Mitchell proposed a WebWatcher recommender model, uses k Nearest Neighbor (kNN) classification method and reinforcement learning. The model acts as a proxy server and uses user's interest terms like keywords and interactively suggests links to user indicates where to go next. The link quality gives the probability that a user will choose the link provided page and interest. Link Quality value for every hyperlink is the average similarity of the k highest ranked keyword sets regarding hyperlink[4].Nizar R. Mabroukeh, C. I. Ezeife [5] and Deshpande Mukund, George Karypis[6] presented Markov model in their paper for prediction. First-order Markov model predicts the next page by looking only at the last user action. In second order Markov model predictions are performed by looking at the last twoactions of user and so the general term is Kth -order Markov model, does prediction by looking the previous K user actions. First-order Markov models are inaccurate in some applications since they cannot look more into the previously accessed web pages while higher-order Markov models gives more accuracy but having state complexity problem.

Mobasher et al. [7] have used the ARM technique in WPP and proposed the frequent item set graph to match an active web user session with frequent item sets and predict the next web page that the user is likely to visit. However, ARM suffers from well-known limitations, including scalability. ARM finds relationships between item sets depending on their co-occurrence in the transactions. This association rule generation can be used to relate pages that are most often referenced together in the sessions. In web usage mining,association rules refer to the sets of web pages that are accessed together with a support value greater than a specified threshold. Here prediction is performed according to the association rules that satisfy certain support and confidence.The scalability problem derived from generating item sets because with the number of item sets it takes exponential time [1]. Mamoun A. Awad and L. Khan presented reduction technique which makes use of domain knowledge for prediction. They have used Artificial Neural Network (ANN) in web navigation incorporated with domain knowledge to remove irrelevant classes. First order frequency matrix represented by N×N is used as domain knowledge where N denotes the no. of pages. Each matrix entry denotes the frequency of two pages visited by user in sequence [8]. F. Khalil, J. Li and H. Wang presented the IMAC model to combine Markov model, Association rules and Clustering. In this model web sessions are clustered then Markov model predictions are generated and association rules are used for prediction if Markov model is unable to make decision [9].

III. ANALYSIS OF PROBLEM

User behavior on the Web can be influenced by individual user characteristics. The specific order Markov Model does the prediction for known patterns only; it is not able to do the prediction for new patterns or patterns which are not available in the training set [1]. ARM causes efficiency and scalability problems. The scalability issue originates from generating item sets which take exponential time with the number of item sets. Thus, it needs to identify user search interests with more accuracy and less prediction time.

IV. PROPOSED WORK

In this section we will describe processes for data preprocessing, All kth Markov and CSB-mine based prediction models to classify the behavior of new users to predict the next web page a user may visit, so the proposed system consist modules as Data preprocessing which performs data cleaning and filtering, users' web sessions construction, Prediction model construction and next page prediction. Figure 1 shows the working flow of the proposed system.

1. Data Preprocessing

A web server log file is considered as a dataset which contains user requests made to the web server. The most popular formats are Common Log Format and extended Common Log Format [10]. The following is an example of single entry in a typical server log.

199.75.81.45 - - [01/Jul/1995:00:00:01 -0500] "GET /history/index.html/ HTTP/1.0" 200 6325

So, given example shows that a user from the IP address 199.75.81.45 successfully requested the page “index.html” on July 01, 1995 at 00:00:01 a.m. The HTTP method the remote user used is “GET”. HTTP Protocol version used is 1.0. HTTP Status code 200 shows OK. The number of bytes returned to the user is 6325.

Some datasets are insufficient, inconsistent and including noise, so data preprocessing is required. Preprocessing is done on the dataset so that necessary data will be available.

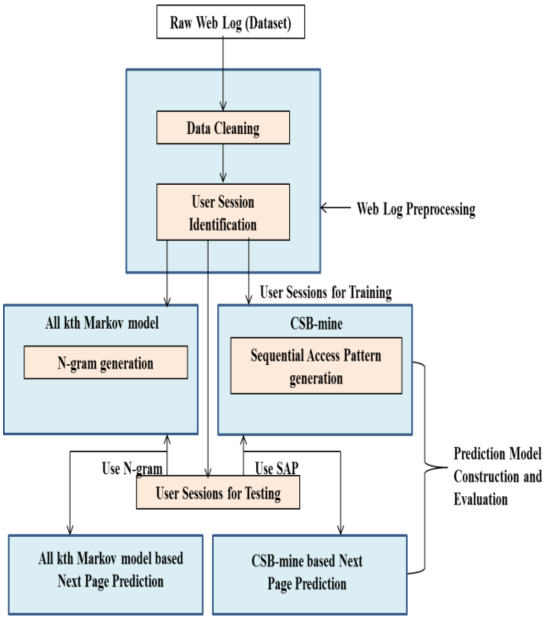
Raw Web Log (Dataset)
Data Cleaning
User Session Identification
Web Log Preprocessing
User Sessions for Training
All kth Markov model
N-gram generation
CSB-mine
Sequential Access Pattern generation
Prediction Model Construction and Evaluation
Use N-gram
Use SAP
User Sessions for Testing
All kth Markov model based Next Page Prediction
CSB-mine based Next Page Prediction

Fig 1: System Work flow

The purpose of data cleaning is to remove irrelevant items, and is important for any type of web log analysis. HTML embeds various types of resources and as HTTP is a connectionless protocol, so a single request for an HTML page results in many log entries for downloading graphics and scripts.According to the purposes of different mining applications, irrelevant records in the web access log will be removed during data cleaning. Since the target of Web Usage Mining is to get the user’s travel patterns, following two kinds of records are unnecessary and should be removed: first, the records of graphics, videos and the format information The records have file name suffixes of GIF, JPEG, CSS, and so on, which can found in the URI field of the every record, and the records with the failed HTTP status code [11]. Web robots are software tools that scan web site content automatically. These tools access the page “robot.txt”, so hosts that request the file isnot considered. Errors’ requests are useless for mining and are removed by checking thestatus of request i.e. the response code. Response codes from 200 to 300 are useful while any other requests are not considered [3].

The goal of user session identification is to divide page accesses of each user into individual sessions. To achieve this task 30 minute timeout period is considered. If the time between consecutive page requests is more than default timeout, a new session is constructed for the respective user. Every session S=<p, q, t, r> is considered as sequence of web pages, so it represents pages p, q, t and r have been accessed by the user in sequence.

2. Model Construction

All Kth Markov model and CSB-mine based prediction model are used for predictions of next web page that a user may visit.

A. All Kth Markov model

Low order Markov models are not able to look far into history of web navigation, so they are not very accurate in many applications. Higher order Markov models have limitation of reduced coverage and specific order of Markov model is unable to predict for a session which is not present in the training set, so the solution is to train different order Markov models i.e. all Kth Markov model. In all Kth Markov model, all orders of Markov model are generated and they are used collectively for performing prediction. Here to build the model sessions are represented as N-grams i.e. sequence of pages surfed by users [1]. Each N-gram accepts page id value to identify specific web page. For example the N-gram <u, q, s, x> represents that user has visited page u first then page q, page s and finally page x. For following sessions different N-grams are calculated associated with their frequency by using sliding window size of 4 to make training examples of the same length.

S1=<u, q, s, v>
S2=<u, q, r>
S3=<s, v, t, p, w>

1-grams for above sessions are <p> 1, <q> 2, <r> 1, <s> 2, <t> 1, <u> 2, <v> 2, <w> 1

2-grams for above sessions are <u, q> 2, <q, s> 1, <s, v> 2, <q, r> 1, <v, t> 1, <t, p> 1, <p, w> 1

3-grams for above sessions are <u, q, s> 1, <q, s, v> 1, <u, q, r> 1, <s, v, t> 1, <v, t, p > 1, <t, p, w> 1

4-grams for above sessions are <u, q, s, v> 1, <s, v, t, p> 1, <v, t, p, w> 1

Following are the steps of the algorithm, where input to the algorithm is user session s, of length K and output p is the next page predicted by the algorithm.

1. predict () function is used to predict the next web page p by consulting K order Markov model m_k

2. If page p is predicted by model m_k then it returns p

3. Else it removes first page id from the session s

4. K= K-1 order of Markov model is used for prediction

5. If (K=0) then it returns with No prediction

6. Continue step 1

7. Stop

For a given user session s=<p, u, q>, all kth model prediction is performed using the third order Markov model, if it fails to predict, then the second order Markov model is used by removing first page id from the session so s=<u, q>. This process continues until the first order Markov model reached. If all orders of the Markov model fail to predict, then only this model fails.

B. CSB-mine

This algorithm is used to generate sequential patterns from sequence database with a user-specified minimum support, i.e. the number of data sequences that contains the pattern, are in turn used for prediction of the next page that a user may visit. Let E be a set of unique access events, which represents web pages accessed by users. A web access sequence $S = e_1e_2\ldots e_n$ ($e_i \in E$) is a sequence of access events and repeat of items is allowed in S. All web access sequences (WAS) in a database represents web user's sessions.In $S = e_1e_2\ldots e_k e_{k+1}\ldots e_n$, $S_{prefix} = e_1e_2\ldots e_k$ is referred as prefix sequence of e_{k+1} in S and$S_{suffix} = e_{k+1}e_{k+2}\ldots e_n$ is called a suffix sequence of e_k in S [12]. The initial conditional sequence base, denoted as Init-CSB, is the set of all WASs in the given database.The conditional sequence base of an event e_i based on prefix sequence S_{prefix}, denoted as CSB(S_c), where $S_c = S_{prefix}+e_i$, is the set of all long suffix sequences of e_i in sequences of a certain dataset [12].

Following are the steps of the algorithm.

1. Provide all WASs and user specified minimum support
2. Prepare an empty Header Table(HT)
3. Identify those events, i.e. Conditional frequent events in CSB (S_c) with support greater than or equal to minimum support given in equation (1), where Init-CSB is the initial Conditional sequence base i.e. set of all WAS

$$\sup(e_i) = \frac{|\{S_j \mid e_i \in S_j, S_j \in CSB\,(S_c)\}|}{|Init-CSB\,|} \geq MinSup \quad (1)$$

4. Remove non frequent events so that HT consists all Conditional frequent events
5. Create event queue, where each item of e_i-queue is the first item labeled e_i in sequence s of CSB(S_c)
6. Construct Sub-Conditional Sequence base and test single sequence for CSB(S_c)

6.1 In CSB (S_c), if it forms a single sequence (if all the i^{th} items in each sequence∈CSB (S_c) are the same event e and total count>=minimum support) then stop with the sequence and will be used as a part of final SAPs

6.2 Otherwise construct Sub CSBs for CSB (S_c) and recursively mine.

7. Return SAP

Table 1: Sample WAS Database

WAS
tpqspr
tptqrpr
qpqup
puqprur
qtprp

For above sample WAS the algorithm works as follows:

Minimum support of 60% the conditional frequent events p=5, q=5, r=4 and t=3 so HT consists p, q, r and t. Each access event is described as event: count, where count represents the number of sequences which contains the event. Header table and Event queue after deleting non frequent events is shown in figure 2.

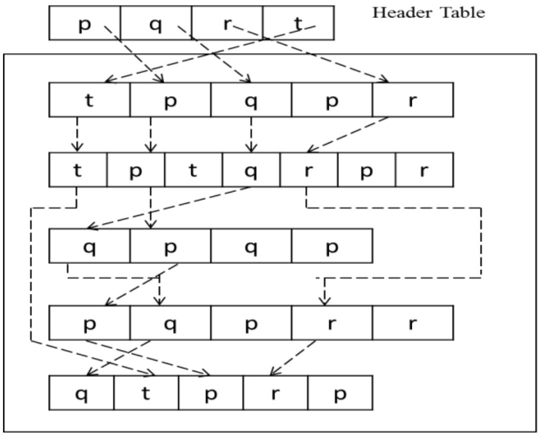

Fig 2: Header table and event queues

CSB (p) contains sequences as qpr, tqrpr, qp, qprr and rp as shown in figure 3, all the sequences cannot form a single sequence.

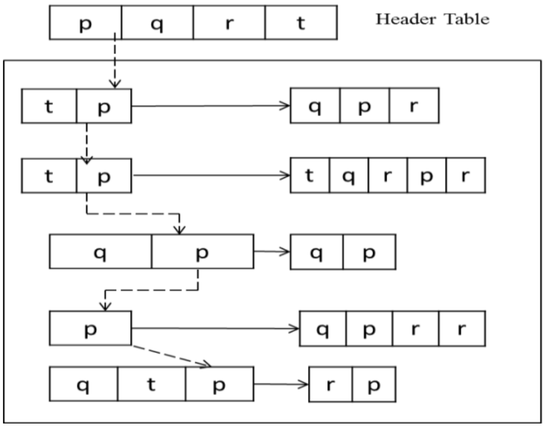

Fig 3: CSB (p) Construction

For CSB (pp) ={r, r, rr}, the sequences can be combined into a single sequence r: 3. This Algorithm recursively mines for CSB (S_C) and generates following SAP of different length. For given user session s= <p, q>, the CSB-mine based model searches for SAP containing current access sequence of user and predicts the next page. If SAP doesn’t contain the given sequence then its suffix sequences are considered for matching by removing first page id until matching or there is no item for removal. From table 2 it will predict pages as p and r.

Table 2: SAP of the Sample WAS

Length	SAP
1	p:5, q:5, r:4, t:3
2	pp:5, pq:4, pr:4, qp:5, qr:4, tp:3, tr:3
3	ppr:3, pqp:4, pqr:3, qpr:4, tpp:3, tpr:3
4	pqpr:3

V. EXPERIMENTAL RESULTS

All experiments have been performed on an Intel Core i5 @ 2.30GHz with 4GB of main memory under Windows 7. The proposed system is implemented in Java using jdk1.8 version and MySQL 5.1 is used for backend for the system. The database consists single table, contains unique pages and their ID number. NASA Kennedy Space Center’s web log which is a common log format of August 1995 is used as a dataset, out of which 100000 web server log entries have been considered for the experiment. The web server log consist all HTTP requests to the NASA Kennedy Space Center WWW server in Florida. From the total number of sessions, 80% sessions are used for building the model and 20% are used for testing.

Number of web log entries:100000
Web log entries after preprocessing: 20956
Number of unique pages: 590
Total Number of user sessions: 7285
Number of Session used for model building: 5829
Number of Session used for model testing: 1456

To build All Kth Markov model, different N-grams, i.e. all 7-grams are calculated by using maximum sliding window of size 7 and for CSB-mine based model, SAP are generated using different support values as shown in table 3. Parameters like running time to build models, memory utilization and prediction time are considered for evaluation.

Table 3: Results of CSB-mine

Support(%)	Time(milliseconds)	Memory (KB)	Number of SAP
0.1	7364	2307	3318
0.15	6692	2015	1140
0.2	5366	3448	661
0.25	4243	3394	454
0.3	3915	3313	364
0.35	3685	3266	285
0.4	3370	2578	239
0.45	3292	2584	202
0.5	3089	2702	179

Table 4 shows the number of SAP generated with different support values and different lengths. Table 5 shows model building time and memory usage comparison between all kth Markov model and CSB-mine based model.

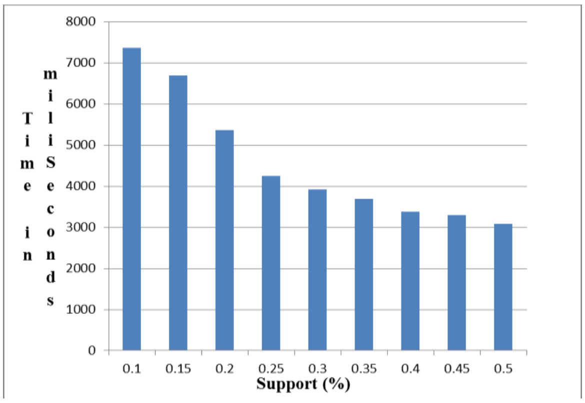

Fig 4: CSB-mine Running time with different support values

Table 4: Number of SAP of Different Length

Support (%)	Length of SAP														Total Number of SAP
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	
0.1	291	1339	1002	427	188	55	8	1	1	1	2	1	1	1	3318
0.15	234	639	239	22	3	1	1	1							1140
0.2	190	371	90	7	1	1	1								661
0.25	152	254	43	2	1	1	1								454
0.3	131	199	30	2	1	1									364
0.35	117	146	19	2	1										285
0.4	102	120	14	2	1										239
0.45	89	103	8	1	1										202
0.5	81	89	8	1											179

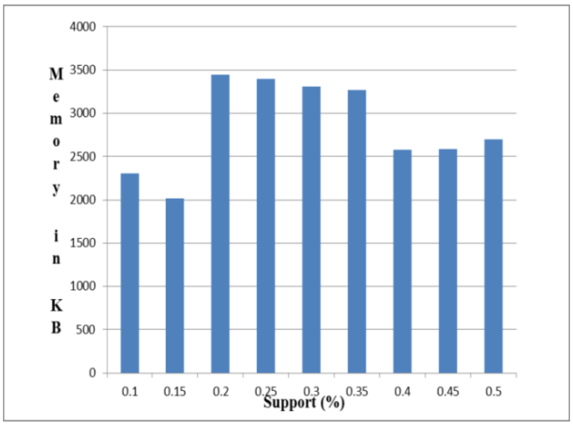

Fig 5: CSB-mine Memory usage with different support values

For prediction using CSB-mine, support value of 0.1 is used and some test sequences are evaluated in the experiment. Maximum top 25 pages have been considered as predictions. If the target page is among top 25 pages, then the prediction is treated as correct otherwise not. Table 6 shows a comparison of prediction time between both the models. Here prediction time is the average of time to predict required for testing sequences.

Table 5: Running time and memory usage for both models

Model	Build Time(milliseconds)	Memory(KB)
All kth Markov	11310	7718
CSB-mine (0.1% Support)	7364	2307

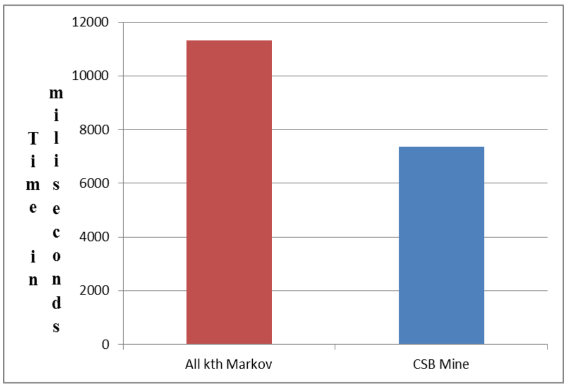

Fig 6: Running time Comparison between All Kth Markov model and CSB-mine based model

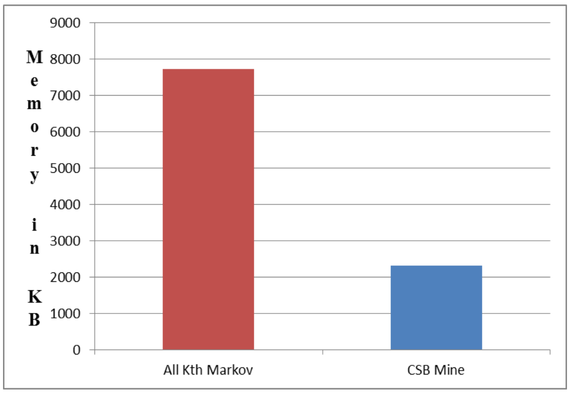

Fig 7: Memory usage comparison between All Kth Markov model and CSB-mine based model

Table 6: Prediction time for both models

Model	Prediction time (milliseconds)
All Kth Markov model	63
CSB-mine (0.1% Support)	30

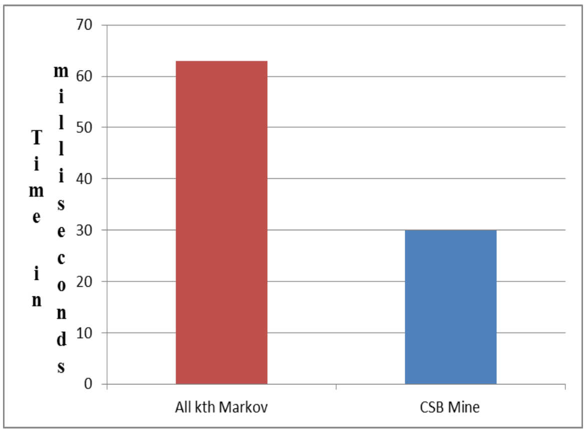

Figure 8: Prediction time Comparison between All Kth Markov model and CSB-mine based model

Prediction accuracy is calculated by Number of correctly predicted testing examples/total number of testing examples.

Model	Prediction accuracy (%)
All kth Markov	41.67
CSB Mine (0.1% Support)	71.67

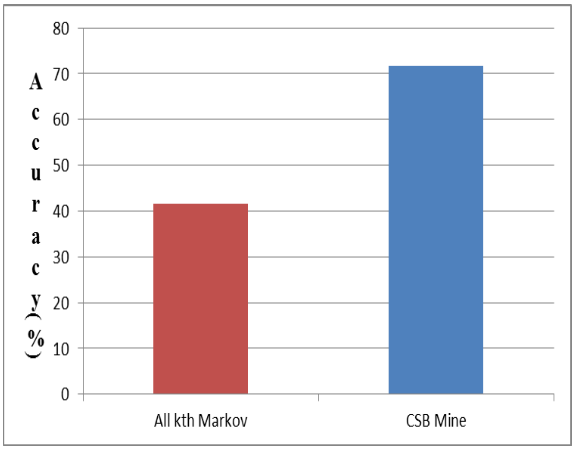

Figure 9: Prediction accuracy comparison between All Kth Markov model and CSB-mine based model

VI. CONCLUSION

Identifying user behavior is very important during web surfing in order to better support users on the Web. This paper focuses on predicting surfing behavior obtained from web server log file. Thus the prediction model can be used for different purposes like web page recommendation, Web page caching, or to find set of Web pages used together to restructure a web site. Experimental results show that to generate different N-grams system takes more time compared to SAP generation using CSB-mine, also it requires less amount of memory compared to all kth Markov model. As minimum support value in CSB-mine decreases, total number of SAP increases. Prediction results show that CSB-mine gives more and correct predictions compared to All Kth Markov model. For all evaluated testing examples, prediction time is more for all kth Markov model compared to CSB-mine, so CSB-mine gives better results in less amount of prediction time.

REFERENCES

[1] Mamoun A. Awad and Issa Khalil, "Prediction of User's Web-Browsing Behavior: Application of Markov Model," *IEEE Trans. on Systems, Man, and Cybernetics.*, vol. 42, no. 4, pp. 1131-1142, August 2012.

[2] Srivastava, T., PrasannaDesikan, and Vipin Kumar, "Web mining–concepts, applications and research directions," *Foundations and Advances in Data Mining. Springer Berlin Heidelberg*, pp. 275-307, 2005.

[3] RajniPamnani and PramilaChawan, "Web Usage Mining: A Research Area in Web Mining," *in Proc. ISCET,* Jun. 2013.

[4] T. Joachims, D. Freitag, and T. Mitchell, "WebWatcher: A tour guide for the World Wide Web," *in Proc. I JCAI,* pp. 770–777, 1997.

[5] Nizar R. Mabroukeh and C. I. Ezeife, " Semantic-rich Markov Models for Web Prefetching," *in Proc. IEEE International Conf. on Data Mining Workshops*, pp. 200-207, Jun. 2009.

[6] Deshpande Mukund and George Karypis, "Selective Markov models for predicting Web page accesses," *ACM Transactions on Internet Technology (TOIT)*, pp. 163-184, 2004.

[7] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa, "Effective personalization based on association rule discovery from Web usage data,"*in Proc. ACM Workshop WIDM,Atlanta, GA*, Nov. 2001.

[8] M. Awad and L. Khan, "Web navigation prediction using multipleevidence combination and domain knowledge,"*IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 37, no. 6, pp. 1054–1062, Nov. 2007.

[9] F. Khalil, J. Li and H. Wang, "An Integrated Model for Next Page Access Prediction,"*International Journal of Knowledge and Web Intelligence*, vol. 1, no.1/ 2, pp. 48-80, August 2009.

[10] http://www.w3.org/Daemon/User/Config/Logging.html

[11] R. Cooley, B. Mobasher, and J. Srivastava, "Data preparation for miningWorld Wide Web browsing patterns,"*J. Knowl. Inf. Syst*., vol. 1, no. 1,pp. 5–32, 1999.

[12] B. Y. Zhou, S. C. Hui and A. C. M. Fong, "Efficient Sequential Access Pattern Mining for Web Recommendations,"*International Journal of Knowledge based and Intelligent Engineering Systems, ACM*, vol. 10, no. 2, pp. 155–168, April 2006.

[13] Internet Traffic Archive. [Online]. Available: http://ita.ee.lbl.gov/html/traces.html