DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.3883

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

WEB PAGE ANTICIPATION SYSTEM USING MARKOV MODEL

Shivangi Sorout
Department of computer science & Application
Maharishi Dayanand University (M.D.U)
Rohtak, Haryana

Dr. Balkishan (Assistant Professor)
Department of computer science & Application
Maharishi Dayanand University (M.D.U)
Rohtak Haryana

Abstract: In the era of digital world, Information can be accessed with the help of Internet on the World Web. Due to the enormous growth of digital information, the limited bandwidth of the network is not utilized in an efficient way. With the help of this "Web Page Anticipation System", we are trying to get rid of this problem. It provides us framework to incorporate the usage of pre-fetching mechanism, Clustering, Markov Model and Prediction Architecture. This framework allows us to anticipate the web page in advance with the help of user's currently accessed web page.

Keywords: com Web page anticipation, Pre-fetchimg Clustering, Markov model, Users sessions

1. INTRODUCTION

Owing to the enormous growth of World Wide Web, congestion and overloading of server occurs. Due to problems incurred by the server in managing large information, latency of communication channel is substantially reduced. Various techniques for latency reduction are web catching, pre-fetching and preopening. Need of this web page anticipation system is required in the era of e-commerce digital world where every transaction is depended only on the efficiency of how fast we are able to accesses the required web page within the particular time slot. Researchers use different kind of techniques comprising Markov Model for next web page anticipation, clustering and prediction Architecture. For implementing this web page anticipation model, navigational behavior of the current users is stored in the web log files.

After the identification of the navigational behavior of current users, Clustering is performed. Clustering is the main ingredient used in the exploratory data mining and commonly used in the statistical data analysis. Main task of clustering is to segregate the group of objects into different groups so that all objects in the one particular group are of similar nature. Scope of clustering is wide and used in the many fields like machine learning, image analysis, information retrieval, bioinformatics, data compression and computer graphics. Clustering models used in the Clustering are connectivity models, centroid based model (k-means algorithm), distribution model expectation-maximization model) and density model.

When the Clusters are formed we have to use the prediction algorithm to predict the next possible states. After that Markov model is applied on the clustering sets. Markov model is a stochastic model which is used for the designing the continuously changing system in variable time period at different time slots. In this model future state is depend not only on the current state but also on the previous states covered. In the Markov model we have to train the model by estimating the transition probability which is denoted by:

$$Aij = P(Q(t+1)=Si \mid Qt=Sj)$$

Where, Aij is the probability of going to the new state Si at time t+1 from state Sj at time t.

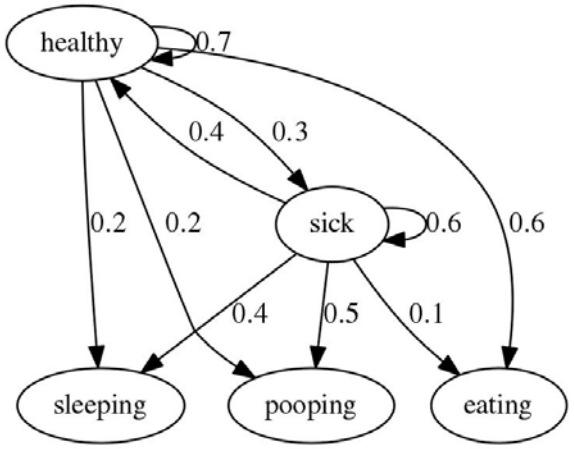

Figure1. Example: Markov Model States and their Transition Probability

The first order Markov model provides us a simple way of to accumulate sequential dependence, however it does not take the aspect of long term memory web navigational behavior. Higher order Markov model are useful for the prediction of navigational path. But with the increase of order of the Markov model subsequently there will be exponential increase in the complexity of state space. In turn we require the huge amount of training data. As the number of states increase, systems which needs to predict fast their prediction accuracy is plummeted to large extent. So the need of the hour is to have that kind of system which predict fast with enhanced accuracy.

2.RELATED WORK

The review of past investigation serves as a guide to the researchers as it avoids duplications in the field. The knowledge of what has already been done in the area of investigation regarding the methods used for data is important. Research in any field implies a step ahead in exploration of the unknown. Any researcher to be able to take this step should be adequately prepared for it. One such preparation is gathering of

knowledge of much has already been done in the given field. A step towards unknown can only be taken after the review of literature and researches done in that area. Any research without such a review is likely to be a building without foundation. Thus, the review of related literature is an indispensable step in research.

Yang et al. (2004) studied different association rule based methods for web request prediction. The association rules for web access prediction involve dealing with too many rules and it is not easy to find a suitable subset of rules to make accurate and reliable predictions. Five different representations of association rules are: Subset rules, Subsequence rules, Latest subsequence rules, Substring rules and Latest substring rules. The author concerned the precision of these five association rules representations using different selection methods, the latest substring rules were proven to have the highest precision with decreased number of rules [1].

Liu et al. (1998) introduced a customized marketing based on the web approach using a combination of clustering and association rules. The author collected information about customers using forms, Web server log files and cookies. It categorized customers according to the information collected. Since k-means clustering algorithm works only with numerical data, the authors used PAM (Partitioning around Medoids) algorithm to cluster data using categorical scales. Then perform association rule techniques on each cluster [2].

Kim et al. (2004) introduced combination of all three models together. It improves the performance of Markov model, sequential association rules, association rules and clustering by combining all these models together. For instance, Markov model is used first. If MM cannot cover an active session or a state, sequential association rules are used. If sequential association rules cannot cover the state, association rules are used. If association rules cannot cover the state, clustering algorithm is applied. The author's work improved recall and it did not improve the Web page prediction accuracy [3].

Vakali et al. (2003) categorized web data clustering into two classes (I) users' sessions-based and (II) link-based. The former uses the web log data and tries to group together a set of users' navigation sessions having similar characteristics. In web log data provide information about activities performed by a user from the moment the user enters a web site to the moment the same user leaves it. The records of users' actions within a web site are stored in a log file. Each record in the log file contains the client's IP address, the date and time the request is received, the requested object and some additional information -such as protocol of request, size of the object etc [4].

Sarukkai et al. (2000) discovered enormous growth in the number of documents in the WWW increases the need for improved link navigation and path analysis models. The author introduced link prediction and path analysis are important problems with a wide range of applications ranging from personalization to web server request prediction. The sheer size of the WWW coupled with the variation in user's navigation patterns makes this a very difficult sequence modeling problem. The author observed Markov chains allow the system to dynamically model the URL access patterns. The Markov chain model can also be used in a generative mode to automatically obtain tours. The Markov transition matrix can be analyzed further using Eigen-vector decomposition to obtain personalized hubs/authorities. The utility of the Markov chain approach is demonstrated in many domains: HTTP request prediction, system driven adaptive web navigation, tour generation, and detection of personalized hubs/authorities from user navigation profiles. The generality and power of Markov chains is a first step towards the application of powerful probabilistic models to web path analysis and link prediction [5].Khalil et al. (2007) introduced concept of prediction of the next page to be accessed by web users. This attracted a large amount of research work lately due to the positive impact of such prediction on different areas of web based applications. Major techniques applied for this intention are Markov model and clustering. Low order Markov models are coupled with low accuracy, high order Markov models are associated with high state space complexity. On the other hand, clustering methods are unsupervised methods, and normally are not used for classification directly. It involves incorporating clustering with low order Markov model techniques. The pre-processed data is divided into meaningful clusters then the clusters are used as training data while performing 2 nd order Markov model techniques. Different distance measures of k-means clustering algorithm are examined in order to find an optimal one. The author revealed that incorporating clustering of web documents according to Web services with low or der Markov model improves the web page prediction accuracy [6].

Deshpande et al. (2001) introduced problem of predicting a user's behavior on a web site has gained importance due to the rapid growth of the World Wide Web and the need to personalize and influences a user's browsing experience. Markov models and its variations found to be well suited for addressing this problem. Different variations of Markov models, it found that higher-order Markov models display high predictive accuracies on web sessions that it can predict. However, higher-order models are also extremely complex due to their large number of states, which increases their space and run-time requirements. The author presented different techniques for intelligently selecting parts of different order Markov models so that the resulting model has a reduced state complexity, while maintaining a high predictive accuracy [7].

Zacharouli et al. (2009) introduced learning algorithms for web page rank prediction, linear regression models and combinations of regression with probabilistic clustering and Principal Components Analysis (PCA). These models learned from time-series data sets and can predict the ranking of a set of web pages in some future time. The algorithm used separate linear regression models. This further extended by applying probabilistic clustering based on the EM algorithm. Clustering allows for the web pages to be grouped together by a mixture of regression models. A different method combined linear regression with PCA so as dependencies between different web pages can be exploited. All the methods evaluated using real data sets obtained from Internet Archive, Wikipedia and Yahoo! ranking lists. It also study the temporal robustness of the prediction framework. Overall the system constitutes a set of tools for high accuracy page rank prediction which can be used for efficient resource management by search engines [8].

Spiliopoulou et al. (1999) investigated web site design is currently based on interests of web site visitors and assumptions about their exact behavior. Concrete knowledge on the way visitors navigate in a web site could prevent disorientation and help owners in placing important information exactly where the visitors look for it. Web utilization miner tool can provide such knowledge. The general problem addressed is given a number of traversed paths

discovers sub-paths with structural or statistical properties of interest. All nodes in a sub-path are of equal Importance. Sub-paths having only some nodes in common be combined into a pattern that shows the desired properties as a whole to capture the ambiguous expressions of this problem. The author described a powerful mining language by which the expert can specify the desired structural and statistical properties of the patterns. To efficiently discover paths which when combined result in such desirable patterns, an innovative technique based on the processing of aggregated sequence several optimization steps are performed to further reduce the mining search space [9].

Mukhopadhyay et al. (2011) studied about pre-fetching models based on decision trees, Markov chains, and path analysis. The author described increase uses of dynamic pages, frequent changes in site structure and user access patterns have limited the efficacy of these static techniques. One of the techniques that are used for improving user latency is Caching and another is Web pre-fetching. Approaches that bank solely on caching offer limited performance improvement because it is difficult for caching to handle the large number of increasingly diverse files. An agent based method is proposed here to cluster related pages into different categories based on the access patterns. Additionally page ranking is used to build up the prediction model at the initial stages when users are yet to invoke any page [10].

Kumar et al. (2011) presented web provides a corpus of design examples unparalleled in human history. Leveraging existing designs to produce new pages is difficult. The author introduced the Bricolage algorithm for automatically transferring design and content between Web pages. Bricolage introduces a novel structured prediction technique that learns to create coherent mappings between pages by training on human-generated exemplars. The produced mappings can then be used to automatically transfer the content from one page into the style and layout of another. The author shown that Bricolage can learn to accurately reproduce human page mappings, and that it provides a general, efficient, and automatic technique for retargeting content between a variety of real web pages [11].

Dutta et al. (2011) studied web page prediction plays an important role by predicting and fetching probable web page of next request in advance, resulting in reducing the user latency. The users surf the internet either by entering URL or search for some topic or through link of same topic. For searching and for link prediction, clustering plays an important role. Web page prediction model give us significant importance to the user's interest using the clustering technique is used for the accumulation of the similar web pages. Similar web pages of same type reside in the same cluster, the cluster containing web pages have the similarity with respect to topic of the session. The clustering algorithms considered are K-means and K-mediods, K is determined by HITS algorithm. Finally, the predicted web pages are stored in form of cellular automata to make the system more memory efficient [12].

Su et al. (2000) studied the rapid development of internet has resulted in more and more multimedia in web content. The author studied due to the limitation in the bandwidth and huge size of the multimedia data, users always suffer from long time waiting. The author describe that to predict the web object or page that the user most likely will view next while the user is viewing the current page, and pre-fetch the content. Then the perceived network latency can be significantly reduced. The author introduced n-gram based model to utilize path profiles of users from very large web log to predict the users' future requests. Model is based on a simple extension of existing point-based models for such predictions, but results show that by sacrificing the applicability somewhat one can gain a great deal in prediction precision. The results can potentially be applied to a wide range of applications on the web, including pre-fetching, enhancement of recommendation systems as well as web caching policies. The experiments based on three realistic web logs have proved the effectiveness of the proposed scheme [13].

Zukerman et al. (2009) used Artificial Intelligence-related techniques to predict user requests. The author implement a learning algorithm such as some variation of Markov chains and use a previous access log in order to train it. This approach also relies on tracking user patterns. Furthermore, it does not handle newly introduced pages, or old pages that have changed substantially. This approach also requires a rather long sequence of clicks from a user to learn his/her access patterns [14].

Safronov et al. (2010) introduced the Page Rank based pre-fetching technique which is a server-side approach and uses the information about the link structure of the pages and the current and past user accesses to drive pre-fetching. The approach is effective for access to web page clusters, is computationally efficient and scalable, and can immediately sense and react to changes in the link structure of web pages. Furthermore, the underlying algorithm uses relatively simple matrix operations and is easily parallelizable, making it suitable for clustered server environments [15].

Padmanabhan et al. (1995) investigated ways of optimizing retrieval latency. Web caching has been recognized as an effective solution to minimize user access latency. A method of called pre-fetching introduced in which clients in collaboration with server pre-fetch web page that the user is likely to access soon, viewing the currently displayed page l. The benefit of pre-fetching is to provide low retrieval latency for users, which can be explained as high hit ratio. This approach reduces web latency by p re-fetching between caching, proxies, and browsers. Web pre-fetching has involved the important issue of log file processing and the determination of user transactions (sessions). It provides various data mining algorithms for the path traversal patterns and how to efficiently mine the access patterns from the web logs [16].

3. PROPOSED MODEL

Web page is the integration of the different web page contains frames, graphics and other information. User cache is used in this model to cache the frequently accessed web page. In the proposed model, we request the certain web page from the server then server will send the URL of that web page to the predictor. After that predictor check the required specific web page, if it is present then predictor gives that web page to the server and server also give required page to the client to meet its requirements. But the predictor while sending the page to the server (in case the predictor is not able to check the requested web page), also give the client's requested web page to the update engine for update and alteration of the data structure. The use of predictor in the process of web page anticipation is that it uses the data structure for storing the web pages. [17]

3.1 **FLOW CHART OF REQUIRED MODEL**

The following flow chart presents the required model and the requisite steps to implement the required Markov model, which in turn helps in web page anticipation. In the proposed model, first step is to give input with the help of preprocessing the web server log files, after which similar web sessions are allocated to appropriate categories. By using clustering, we decide the number of clusters and among these clusters web sessions are partitioned. The process of clustering gives us the clustered data which is used for the Markov model approach.

When the Markov model is applied we have decide the prediction algorithm and then apply the hidden Markov model in prediction algorithm. After the determination of the prediction algorithm the next web page for user access is available as output.

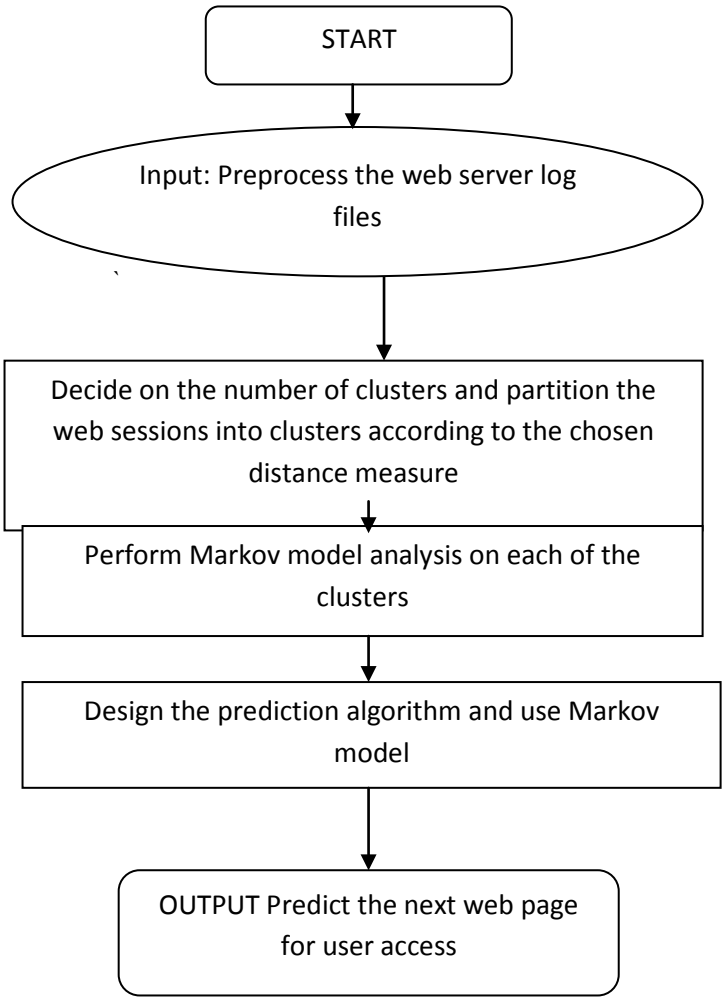

Figure2. flow chart for web page anticipation

1. STEP1:
web server log files are preprocessed in a way similar web sessions are allocated to appropriate categories.

2. STEP2:
Convert the web pages into numeric form and store in web.dat file and then determine the number of clusters and then division of web sessions into clusters is done using clustering tool based on FCM algorithm on Matlab.

The FCM Algorithm uses the k-means clustering to choose the number of clusters k. In this clusters centers U1 to Uk are decided after that we could pick k data points and set cluster centers to these points or it could randomly assign points to clusters and take means of clusters. For each data point, we decide the cluster center it is closest to and assign the data point to this cluster.

After that with the help of findcluster command on the command prompt on the matlab, the web.dat file is loaded on Matlab.

3. **STEP3**:
Perform Markov model analysis on each of the clusters and then make squared Transition probability matrix and squared Emission Probability Matrix.
Making of squared Transition Probability Matrix (Rows=Columns=Total number of unique web pages=9) TRANSITION (I,J) is the probability of transition from state I to state J
Making of squared Emission Probability Matrix (Rows=Columns=Total number of unique web pages=9) EMISSION(K,L) is the probability of transition from state K to L.

4. STEP4:
Design the prediction algorithm and use hidden Markov Model approach. This algorithm gives us information about the next page with the help of user's currently accessed web page.

ALGORITHM:-

1. Set Tr=Transition square matrix, Tr(I,J)=Probability of transition from state I to state J. [Initializes Tr]
2. Set E=Transition square matrix, Tr(K,L)=Probability of transition from state K to state L. [Initialize E]
3. Set seq=sequence of user's accessed web page. [Initialize seq]
4. numStates= number of states and size=size of any column Tr.
   numStates=size
5. L=length of seq
6. Repeat for count 1 To L
7. Repeat for state 1 to numStates
8. Set bestVal=0 [initialize bestVal]
9. Set bestPTR=0 [initialize bestPTR]
10. Repeat for inner 1to numstaes

Val=Tr[innerstate]

If val> bestVal

bestVal=val

bestPTR=val

bestPTR= inner

[End of if structure]

[End of step 10 inner loop]

11. PTR[state,count]=bestPTR

12. v[state]=E[state,seq[count]]+ bestval
[End of 7 inner loop]

13 ..vOld=v

14. P=max[v] max is maximum value

15. finalState= max[v]

16. currentState[count]=finalState

[End of step 6 outer loop]

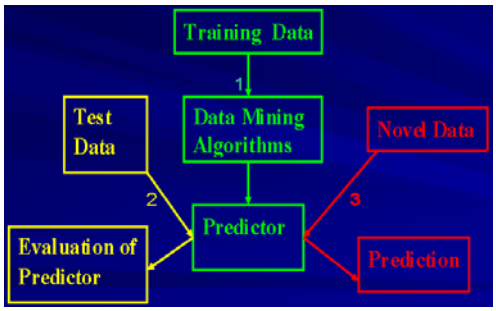

Figure 3. Prediction Architecture for training Data

5. STEP 5:

If we execute this prediction algorithm on the matlab then this will give a next probable webpage for user's currently accesses webpage.

4. RESULTS

Merging web pages by web services according to their function in turn reduces the number of unique pages. The sessions were divided into varying clusters using k-means algorithm and cosine function measure. For each cluster, the categories were expanded back to their original form in the data set. This process is performed using a simple program that seeks and displays the data related to each category. Markov model implementation was carried out for the clusters. Markov model accuracy was calculated accordingly. Then, using the test set, each transaction was considered as a new point and distance measures were calculated in order to define the cluster that the point belongs to. Markov model prediction accuracy was computed considering the transaction as a test set and only the cluster that the transaction belongs to as a training set. Prediction accuracy results were achieved using the maximum likelihood based on conditional probabilities.

All implementations were carried out using MATLAB. The Markov model accuracy was calculated using a 10-fold cross validation. Results collected are user's currently accessed web page and Next web page for user's currently accessed web page. The reported accuracy is the how many Next web pages are user's actually accessed web pages after user's currently accessed web pages. Markov model accuracy using clusters based on Cosine distance measures with k = 4.

All clustering runs have performed on a desktop PC with a Pentium IV Intel processor running at 2 GHz with 2 GB of RAM and 100 GB of hard disk memory. . The runtime of the k-means algorithm, regardless of the distance measure used, is equivalent to *O(nkl)*, *n* is the number of items, *k* is the number of clusters and *l* is the number of iterations taken by the algorithm to converge. For experiments, *n* and *k* are fixed, the algorithm has a linear time complexity in terms of the size of the data set. The k-means algorithm has a $O(k + n)$ space complexity. This is because it requires space to store the data matrix. It is feasible to store the data matrix in a secondary memory and then the space complexity will become *O(k)*. k-means algorithm is more time and space efficient than hierarchical clustering algorithms with $O(n^2 \log n)$ time complexity and $O(n^2)$ space complexity.

Figure 4. Markov Model accuracy

5. CONCLUSION

This paper describes the overall prediction accuracy by grouping the data set sessions into clusters and reduces the web latency time. The web pages in the user sessions are segregated into categories according to Web services that are functionally meaningful. Then k-means clustering algorithm is implemented using the most appropriate number of clusters. The Prediction algorithm for the determination of user's next probable page with the help of user's currently accessed page is applied on the number of clusters. The results gives us the accuracy of the next page access prediction by implementing the k-means clustering algorithm on the data set decided previously. More accuracy describes most accurate web page predicted with the help of prediction algorithm implementation as user want next time. Now user do not need to request the web page as user wants, because it has been available before the time user want it next time. In this way we are able to utilize the minimum bandwidth of the user and the use of pre-fetching and clustering reduces the request made by user and reduces request time. The prediction accuracy achieved is an improvement over the problems occurred to the user at the time of accessing web information on the Internet.

6. FUTURE WORK

This paper introduced the Prediction algorithm for automatically transferring Web pages. It demonstrated that it can learn to closely reproduce human mappings, and it take a one step towards a powerful new paradigm for instance- web based design and opens up exciting areas for research. At present, the algorithm employs only about thirty simple visual and semantic features. Expanding this set to include more complex and sophisticated properties, such as those based on computer vision, will likely improve the robustness of the machine learning. Additionally, this implementation cannot handle idiosyncrasies of modern HTML. Extending Prediction to these technologies remains future work.

7.REFERENCES

[1] Jia Yang, J. Zhang, and K. Beach, "A Survey of Web Caching Schemes for the Internet", ACM SIGCOMM, 2004.

[2] M. Liu , M. J unchang, and G . Zhimin, " Finding Shared Fragments in Large Collection of Web Pages for Fragment-based Web Caching", Second IEEE International Symposium on Network Computing and Applications (NCA'06), 1998.

[3] D. Kim, N. Adam, V. Alturi, M. Bieber, and Y. Yesha, "A Click Stream b ased Collaborative F iltering P ersonalization Model: Towards a b etter p erformance". WIDM ' 04, pages 8 8–95, 2004.

[4] W. V akali, S . Yu, and D . Cai, "Improving ps eudo- relevance Feedback i n W eb I nformation R etrieval using W eb Page Segmentation", I n P roceedings of t he Twelfth I nternational World W ide W eb C onference, W WW2003, pp. 11-18, Budapest, Hungary, May 20-24, 2003.

[5] R. Sarukkai, "Link P rediction a nd P ath Analysis u sing Markov C hains", 9th I nternational W WW C onference, Amsterdam, pages 377–386, 2000.

[6] Faten K halil, "Integrating Markov M odel with C lustering f or Predicting W eb Page A ccesses", W eb de velopment a nd mining, 2007.

[7] Mukund D eshpande, "Selective Markov Models for Predicting Web-Page A ccesses", IEEE/WIC/ACM I nternational Conference on Web Intelligence, 2001.

[8] Polyxeni Z acharouli, "Web P age Rank P rediction with P CA and EM Clustering", Web Engineering and Web mining, 2009.

[9] Myra Spiliopoulou, "A D ata Miner a nalyzing th e N avigational Behaviour of Web Users", Data Mining, 1999

.[10] Debajyoti M ukhopadhyay, "Hybrid Web P age P rediction Model f or Predicting U ser's Next A ccess", Information Technology Journal, 9: 774-781, 2011

.[11] R anjitha K umar, "Bricolage: A S tructured-Prediction A lgorithm for E xample-Based W eb Design", W eb development and D ata mining, 2011.

[12] Ruma D utta, " Clustering-based web pa ge p rediction", Int. J . Knowledge and Web Intelligence, Vol. 2, No. 4, 2011.

[13] Zhong Su, "A P rediction System for M ultimedia Pre-fetching in Internet", Microsoft Research China in Beijing, 2000.

[14] Zukerman, S. Y ang, J. Z hang a nd S . T sai, " An a utomatic Semantic- Segment Detection M ethod i n the HTML Language", S ervices Computing, 2008. S CC a pos;08 IEEE International Conference on Volume 1, Issue , 2009.

[15] Victor Y. Safronov, "Strong Cache Consistency on World Wide Web", 3r d I nternational Conference o n Advanced C omputer Theory and Engineering(ICACTE), 2010.

[16] Padmanabhan, "Using P redictive P refetching t o I mprove World Wide Web L atency", C OMPUTER COMMUNICATION, 1996.

[17] http://www.ijcmc.com/docs/papers/May2015/V415KJ0 3.pdf

