Enhancing Proxy Server Cache Management using Log Analysis and Recommendations

Madhubala Chaurasia
Department of IT
Medi-Caps Institute Technology and Management, Indore (M.P.)

C.S. Satsangi, Ph.D.
Department of IT
Medi-Caps Institute of Technology and Management, Indore (M.P.)

ABSTRACT
Now a day's web based applications are growing rapidly due to this the network performance is affected significantly. Thus a performance improvement technique is required by which the application speed is maintained and delivers the high performance web pages. Thus pre-fetching techniques are applied. There are various kinds of pre-fetching techniques are available among them a promising data model is find in [1]. In this technique the proxy log data is consumed for performing navigation pattern analysis. Thus this model incorporates a K-mean algorithm for cluster log data and then the Apriori algorithm is applied to find the frequent pattern rules. Using these rules the system recommends the possible user pages for prefetching. In order to enhance the performance of the traditional model two different techniques are implemented and compared with the traditional model. First utilize the Bayesian classification technique for analyzing pattern and in second method the ID3 decision tree algorithm for analyzing patterns. The comparison of the both the techniques are performed in terms of memory used, time consumption, accuracy and error rate. According to the obtained results the proposed predictive system offers high performance results as compared to the traditional data model.

Keywords
web usage mining, web log, prefetching , ID3, K-mean, Bayesian classifier.

1. INTRODUCTION
Web mining is an application of data mining technique to automatically discover pattern and extract information from the web. This area is most popular among today research .The web define as universal body of virtual data available for processing. Web mining fetch information from web data or services and applied to semi structure or unstructured data [2].

Web content mining is useful to search web pages through content. Content mining is the scanning, mining, extraction and integration of text, pictures and graphs of a Web page to determine the data or information of the content.

Web Usage Mining is technique to predict the user behavior while they are interacting with the web. Web usage mining, discover user navigation path from web data and tries to discover the useful information from the web while interacting with user.

Web structure mining is the process of categorizing the Web pages and give the information of different Web sites. Web structure mining Discover useful structure information from the web pages and it is good for the navigation purpose.

With the growth of internet use of website across the world has been increase rapidly and used for communication purpose. Network congestion has become the biggest problem due to the massive use of World Wide Web, a quick web search and response to user query is essential for effective utilization of web. More suitable software is pre-fetching to solve the problem like network congestion, low bandwidth, bandwidth underutilization and propagation delay. The web perfecting is one of the technique propose to reduce user perceived latencies in the World Wide Web. The special locality shown by user's accesses make it possible to predict future access based on the previous one and pre load those into the cache. The perfecting objects are store in the cache to reduce the latency time. Perfecting technique has two main components. Prediction engine run a prediction algorithm to predict the next user request and perfecting engine is used to decide prefetching them or not depending upon some condition [3].

2. BACKGROUND
Weblog is a discussion or information site published on the web that consist series of entry and it updated frequently with the new information about particular topics. A Web Log mining is the process of extracting data to discover the usage patterns from web server logs and web browser activity tracking systems to understand and better serve the needs of users by reorganization of website. Web logs have been considered as the best tools for understanding customers as these helps on knowing how the customers find your website and why they are looking for it. We can analyze web logs to know the access patterns of users, which provide information regarding most frequently visited web pages of our website [4].

Log files Classified into three categories depending on the location of their storage.

• **Web Server Log Files:** These log files resides in web server and notes activity of the user browsing website.

• **Web Proxy Server Log Files:** These log files contains information about the proxy server from which user request came to the web server.

• **Client browser Log Files:** These log files resides in client's browser and to store them special software are used.

2.1 Log Files Parameters
Log files contain various parameters which are very help full to recognizing user browsing patterns.

• **User Name:** Identifies the user who has visited the website and identification of the user mostly would be the IP address that is assigned by the Internet Service provider.

• **Visiting Path:** It is the path visited by user on the website. This may be by using the URL directly or by clicking on a link.

• **Path Traversed:** It is the path travel by user within the website through various links.

• **Time Stamp:** It is the time spent by user on each page and is normally known as session.

• **Page Last Visited:** It is the page last visited by the user while leaving the website.

• **User Agent:** It is the browser that uses by user to send the request to the server.

• **URL:** It is the resource that is accessed by the user and it may be of any format like HTM page, CGI program or a script etc.

• **Request Type:** the GET or POST method that is used by the user to send the request to the server.

3. LITERATURE REVIEW

This section includes the recent contributions on the web pre-fetching techniques development and their improvements.

Ms.Veena [5]*et.al* Suggest the use of FP growth to finding frequent page efficiently without candidate set generation and weighted rule mining concept apply relative weight over each transaction and Markov model is used to assign that relative weight over their relative position in transaction probability chain matrix for fast and frequent web pre fetching to improved the user hit ratio of web page and provide moderate time and space complexity.

Kushvant kaur[6]*et.al* suggest a hybrid technique that combines the web caching and the web perfecting technique and doubles the performance of proxy server as compared to single caching. Pre-fetching the predicted web objects into the proxy server cache can increase cache hit-ratio, and results in fast retrieval of web page. They also suggest the concept of optimized hybrid approach, which bring preference list from Dynamic technique into Domain top technique. It reduces the ideal time of the existing network and makes traffic almost constant.

Akshay Kansara[7]*et.al* suggest navigation pattern prediction system. The main objective of the proposed system is to predict user navigation behavior using information from a Classification process that identifies potential users from web log data and a clustering process that groups users with similar interesting habits and Using the results of classification and clustering, predict future user request and identify potential users. A clustering algorithm was used to discover the navigation pattern.

P. jomsri[8] suggests data mining technique for improving proxy server performance based on the relation of web content. In the association rule mining, the support and confidence value were used to determine the web relation and only one-to-one association rule is considers such as time, main group web, sub group web Domain name can help to predict web page and improve the performance of the proxy server. Particularly, the useful of prediction web applications can increase Hit Rate of proxy server

Nanhay Singh[9] *et.al* suggest a framework to improve the proxy server performance using web usage mining and perfecting technique. It use k-mean algorithm to cluster the user according to their access behavior then use Apriori algorithm to generate rules for perfecting the pages. These cluster based approach applied on web proxy log data to test the result and improve the hit ratio and byte ratio.

Naveed Ahmad[10]*et.al* suggest the sequential rank selection algorithm that select only one web page of a web site for prefetching purpose and the pages locally available to a user or group of users by utilizing bandwidth of the network and consume less memory space of user. It reduces the user latency due to the efficient prediction of web pages by the Sequential Rank Based Selection algorithm from the cluster.

Mahesh Manchanda[11]*et.al* suggest application of web log mining to obtain web-document access patterns of closely related pages based on the analysis of the request from the proxy server log file. It use LRU technique using this approach, we identify frequently requested web pages by the users and integrate the pre-fetching scheme with the web caching to achieve performance improvement for the proxy server cache. There was performance improvement in terms of the Hit Ratio and the Byte Hit Ratio.

4. PRAPOSED WORK

4.1 proposed Algorithm ID3

The system recommended URL for which user can hit this URL should be pre-fetched and store for reducing time and effort of any search system or application. Our system firstly find the hit probability for each unique URL in user web access log data .Then 60% of URL which have maximum hit probability should be fetched and then this URL become input for the ID3 decision tree with time and agent information. The ID3 can generate the rules that show some URL can be accessed mostly on some mention time on the mentioned agent or web browser. This kind of rules can also yield the information in prospects of URL perfecting with particular timing and agent information.

Entropy

In order to define information gain exactly, we first require discussing entropy. Entropy is a measure of the amount of uncertainty in data. Let's assume, that the decision tree classifies no of instances into two different categories, we'll call them p (positive+) and p (negative-).

Given a set A, containing these positive (+) and negative(-) targets, the entropy of A is calculated by this equation:

Entropy (A) = - P (positive+) $\log_2$P (positive+) –P (negative-) $\log_2$P (negative-)

P (positive+): proportion of positive examples in A, P (negative-): proportion of negative examples in A.

Information Gain

Information gain measure the difference of entropy from before to after the set is split on an attribute and minimize the uncertainty after splitting set on attribute. Information gain is calculated for remaining attribute. The attribute with the high information gain is used to split the set on this iteration.

The information gain, Gain (A, B) of an attribute B,

$$Gain(A,B) = Entropy(A) - \sum_{n=1}^{v} \frac{A_v}{A} XEntropy(A_v)$$

We can use this equation of gain to rank attributes and to build decision trees where each node is locate the attribute with have high information gain among the attributes not considered in the path from the root.

4.2 Bayesian Classifier

The Naive Baye's classification algorithmic rule is a probabilistic classifier. It is based on probability models that incorporate robust independence assumptions. The independence assumptions usually don't have an effect on reality. So they're thought of as naive. You can derive probability models by using Baye's theorem (proposed by Thomas Baye's). Based on the nature of the probability model, you'll train the Naive Baye's algorithm program in a very supervised learning setting. In straightforward terms, a naive Baye's classifier assumes that there is no dependency of a specific feature is presence or absence of the other feature, given the category variable. There are two types of probability as follows:

- Posterior Probability [P (H/X)]
- Prior Probability [P (H)]

The prior probability is a probability that is obtain from previous information and the posterior probability is obtain from new information that is experience later.

Where, X is data tuple and H is some hypothesis. According to Baye's Theorem

$$P\left(\frac{H}{X}\right) = \frac{P\left(\frac{X}{H}\right)P(H)}{P(X)}$$

4.3 Architecture:

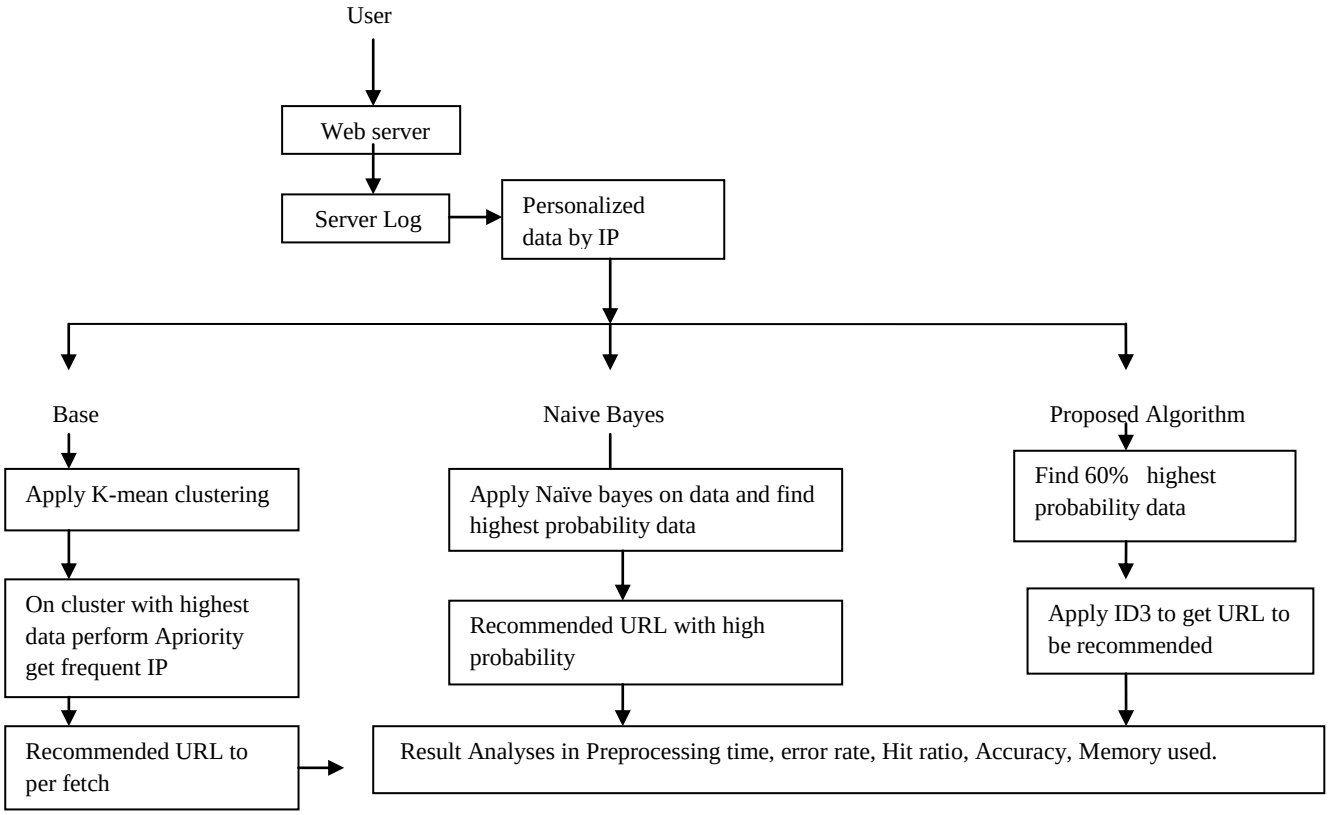

Figure.1 Proposed system architecture

4.4 Proposed Algorithm

Table.1 proposed Algorithm

Input: Examples, Target_Attribute, Attributes
Output: Decision Tree
Process: • Create a root node for the tree • If all Attributes are positive, Return the single-node tree Root, with label = +. • If all Attributes are negative, Return the single-node tree Root, with label = -. • If number of expecting attributes is empty, then Return the single node tree Root, with label = most general value of the target attribute in the examples. • Otherwise Begin ○ A = The Attribute that best classifies instances. ○ Decision Tree attribute for Root = A. ○ For each possible value, v_i, of A, ▪ Add a novel tree branch below Root, equivalent to the test A = v_i. ▪ Let Examples(v_i) be the subset of instances that have the value v_i for A ▪ If Examples(v_i) is empty • Then below this novel branch insert a leaf node with label = most common target value in the examples

- Else below this new branch add the sub-tree ID3 (illustrations(v_i), Target_Attribute, Attributes – {A})
- End
- Return Root

5. RESULT ANALYSIS

In order to demonstrate performance of the proposed algorithm some experiment were performed. The result is analyzed in terms of –

Preprocessing Time- The preprocessing time indicate the time taken in extraction of instances with high probability.

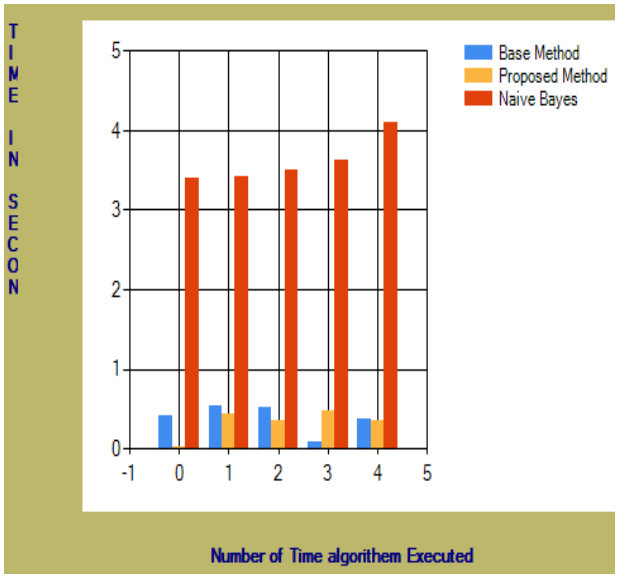

Figure2. Preprocessing Time

Time Taken- The time consumption of the system is also termed as the time complexity. Time consumption of the system denotes the amount of time required to processes the algorithm.

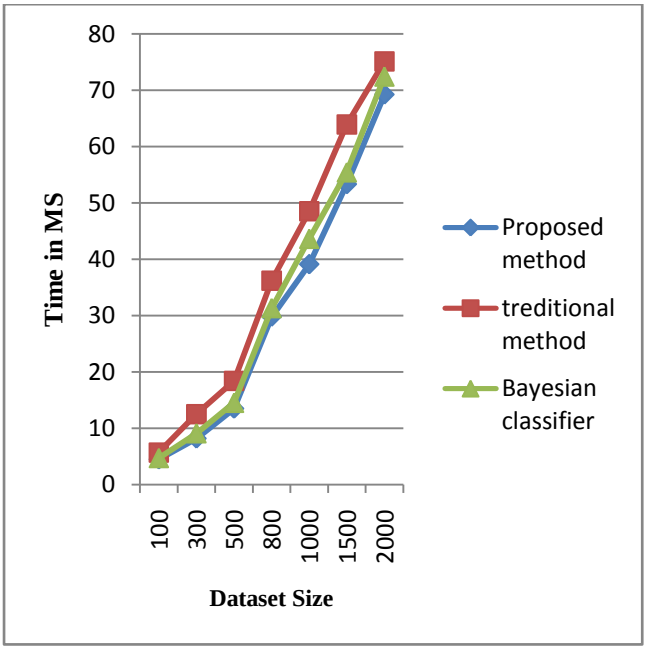

Figure 3 Times Taken

Accuracy- In the predictive systems the accuracy of the system is given as the amount of data that correctly recognized during classification of the input data. the accuracy of the system can be evaluated using the following formula:

$$accuracy = \frac{total\ correctly\ classified\ samples}{total\ samples\ provided\ as\ input} X100$$

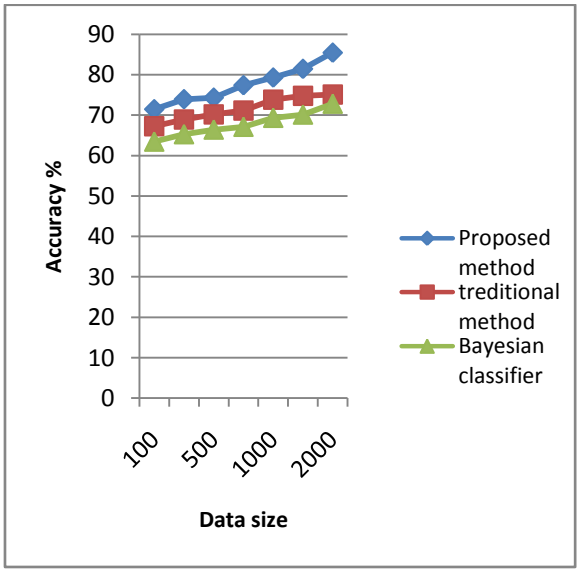

Figure 4 Accuracy

Memory Used- Memory consumption is sometimes also termed as the space complexity. The space complexity demonstrates the amount of main memory required to successfully execute the algorithm.

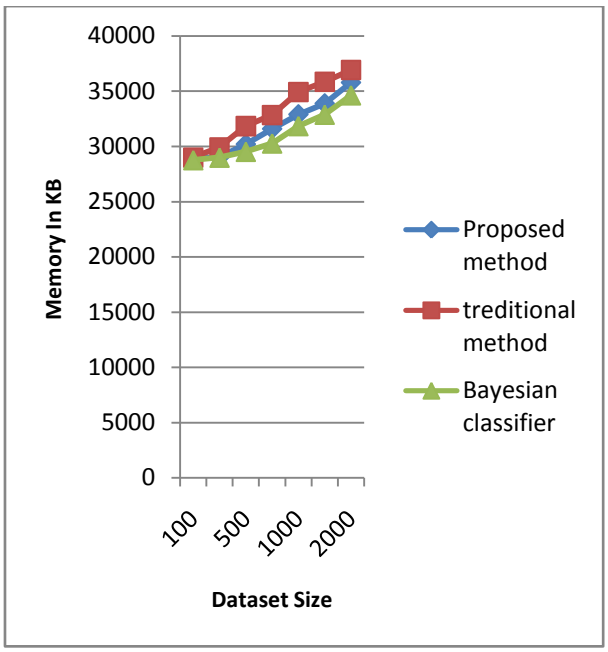

Figure 5 memory used

Error rate- Error rate of the predictive algorithm demonstrate the amount of data which is not correctly recognized using the algorithm. That can be computed using the given formula:

$$error\ rate = 100 - accuracy$$

Or

$$error\ rate = \frac{total\ incorrectly\ classified\ samples}{total\ samples\ provided} X100$$

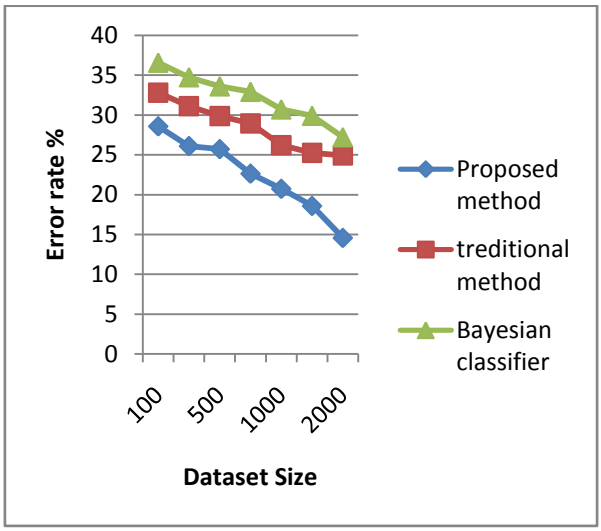

Figure 6 Error rate

Hit Ratio-For calculation of this parameter we have create some data to make test set and some part of data become train set from the overall web access log data. The train data is for training of the system and train set is used for calculating hit ratio. Which shows that if a user hit for a URL then is it in our cache or not.

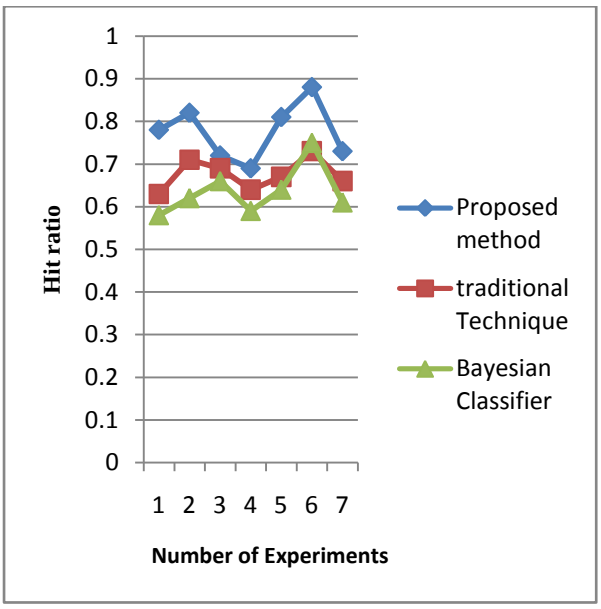

Figure.7 Hit Ratio

6. CONCLUSION AND FUTURE WORK

The main motive of the proposed work is to investigate the web browsing performance enhancement techniques study in addition of that improving the performance of a traditional pre-fetching technique. Therefore a number of research articles and paper are studies. There are some issues and challenges are addressed, first the performance accuracy and second is computational cost. Thus a new data modal is required which offer high accuracy during prediction with less computational cost. The proposed data model includes a hybrid approach of data analysis and prediction. Therefore three different algorithms are applied in a sequence. First the input web access log of the proxy server is pre-processed and stored in an intermediate storage. Where a K-means clustering algorithm is applied and IP address based clusters are produces. This user based clustered data is used to develop ID3 based decision rules. Than after for the current user navigation sequences based next URLs for pre-fetching is performed in this step KNN (K-nearest neighbor) algorithm is applied for prediction. The performance evaluation of the system is performed in terms of accuracy, error rate, memory consumption and time consumption.

Table.2 Performance summary

S. No	Parameters	Proposed method	Traditional method	Bayesian classifier
1	Accuracy	High	Low	Low
2	Error rate	Low	High	High
3	Memory consumption	Average	High	Low
4	Time consumption	Low	High	Average
5	Hit Ratio	High	Average	Low

According to the evaluated results the performance of the proposed data model is adoptable due to high accurate predictive results, reducing error rate and low time and space complexity.

In near future the performance of the proposed predictive system is enhanced more by applying the other classification techniques

7. ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their helpful suggestions and contribution to improve the quality of the paper.

8. REFERENCES

[1] Nanhay Singh, Arvind Panwar, and Ram Shringar Raw, "Enhancing the Performance of Web Proxy Server through Cluster Based Perfecting Techniques", 978-1-4673-6217-7/13/$31.00 c 2013 IEEE.

[2] R. Suguna, D. Sharmila, "Clustering Web Log Files – A Review" International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 4, April – 2013 ISSN: 2278-0181.

[3] Nanhay Singh, Achin Jain1, Ram Shringar Raw "Comparison Analysis Of Web Usage Mining Using Pattern Recantation Technique" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.4, July 2013 .

[4] Hamid Rastegari and Siti Mariyam Shamsuddin, "Web Search Personalization Based on Browsing History by Artificial Immune System", Int. J. Advance. Soft Compute. Appl., Vol. 2, No. 3, November 2010 ISSN 2074-8523; Copyright © ICSRS Publication, 2010.

[5] Ms. Veena Singh Bhadauriya1, Dr. Bhupesh Gour2, Dr. Asif Ullah Khan" A Weighted Markov Model for Web Prefetching to Improve User Interface over Internet" Int.

J. Advanced Networking and Applications Volume: 05, Issue: 03, Pages:1962-1967 (2013) ISSN : 0975-0290.

[6] Kushwant Kaur, Prof. Kanwalvir Singh Dhindsa" Hybrid Approach for Improvement of Web page. Response Time" Kushwant Ka et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014, 6755-6759.

[7] Akshay Kansara1, Swati PatelI mproved Approach to Predict user Future Sessions using Classification and Clustering" International Journal of Science and Research (IJSR), India Online ISSN: 2319-7064.

[8] P.jomsari"improveing the performance of proxy server by using data mining technique"world Academy of science,Engineering and technology vol:7 2013-07-24.

[9] Monti Babulal Pal1, Dr. Dinesh C. Jain" Enhancing the Web Pre-Fetching at Proxy Server using Clustering" Engineering Universe for Scientific Research and Management Vol.6 Issue 3 March 2014.

[10] Naveed Ahmad, Owais Malik, Mahmood ul Hassan, Muhammad Shuaib Qureshi, Asim Munir, "Reducing User Latency in Web Prefetching Using Integrated Techniques", 978-1-61284-941-6/11/$26.00 ©2011 IEEE.

[11] Mahesh Manchanda,Dr. Neena Gupta, "Make Web Page Instant: By Integrating Web-Cache and Web-Prefetching", Conference on Advances in Communication and Control Systems 2013 (CAC2S 2013).