SOM Improved Neural Network Approach For Next Page Prediction

Vidushi[1], Dr. Yashpal Singh[2]

1Research Scholar, Department of Computer Science and Engineering,
Ganga Institute Of Technology & Management, Kablana

2 Associate Professor(CSE Dept.)
Ganga Institute Of Technology & Management, Kablana

Abstract-

The World Wide Web can be considered as a large distributed information system that provides access to shared data objects. As one of the most popular applications currently running on the Internet, The WWW has become a huge, diverse, and dynamic information reservoir accessed by people with different backgrounds and interests. On the Web, access information is generally collected by Web servers and recorded in the access logs. Web mining and user modeling are the techniques that make use of these access data, discover the surfer's browsing patterns, and improve the efficiency of Web surfing. The World Wide Web is of an exponential growth in size, which results in network congestion and server overloading. Also, the WWW has documents that are of diverse nature and so everyone can find information according to their liking. But, this scorching rate of growth has put a heavy load on the Internet communication channels. This situation is likely to continue in the foreseeable future, as more and more information services move onto web. The result of all this is increased access latency for the users.[1]

1.Introduction-

1.1Requirement of caching in web

A Web cache is a mechanism for the temporary storage (caching) of Web documents, such as HTML pages and images, to reduce bandwidth usage, server load, and perceived lag. A Web cache stores copies of documents passing through it; subsequent requests may be satisfied from the cache if certain conditions are met [2].A Cache is a component that transparently stores data so that future requests for that data can be served faster. The data that is stored within a cache might be values that have been computed earlier or duplicates of original values that are stored elsewhere. If requested data is contained in the cache (cache hit), this request can be served by simply reading the cache, which is comparatively faster. Otherwise (caches miss), the data has to be recomputed or fetched from its original storage location, which is comparatively slower. Hence, the more requests can be served from the cache the faster the overall system performance is [2].To be cost efficient and to enable an efficient use of data, caches are relatively small. Nevertheless, caches have proven themselves in many areas of computing because access patterns in typical computer applications have locality of reference. References exhibit temporal locality if data is requested again that has been recently requested already. References exhibit spatial locality if data is requested that is physically stored close to data that has been requested already.

In simple terms, Web caching is a technology that can significantly enhance end-user's Web browsing experience and, at the same time, save bandwidth for service providers. In detail, a Web cache is a temporary storage place for data content requested from the Internet. After an original request for data has been successfully fulfilled, and that data has been stored in the cache, further requests for those files (e.g., HTML pages, images) results in the information being returned from the cache, if certain conditions are met, rather than the original location. Web Caching is the widely used technique, used by Internet Service Providers (ISPs) all around the world, to save bandwidth and to improve user response time. In short, Web caching temporarily stores Web objects – HTTP and FTP data – flowing into ISP's network. This is not an entirely new invention in that Caching is an integral part of computer architecture, for example CPU cache speeds up an access to main memory,

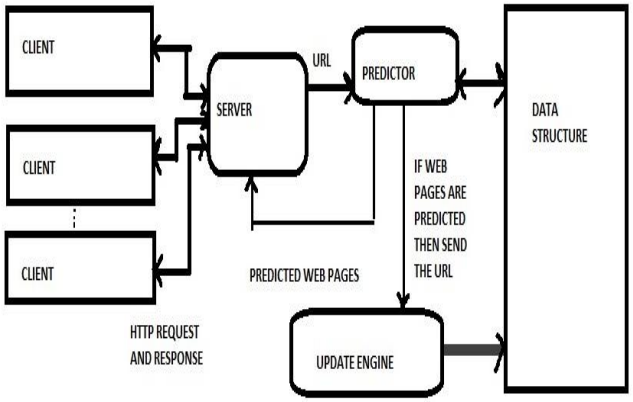
ISSN: 2278 – 1323

file system cache stores commonly requested blocks for faster access and so on[2].

1.2Web page prediction

Predictive caching is the speculative retrieval of a resource into a cache based on user access log; in the anticipation that it can be served from cache in the future [Padmanbhan 1995] leading to improvement in web server performance. Pre-fetching attempts to transfer data to the cache before it is asked for, thus lowering the cache misses even further. Pre-fetching techniques can only be useful if they can predict accesses with reasonable accuracy and if they do not represent a significant computational load at the server. Note that pre-fetching files that will not be requested not only wastes useful space in the cache but also results in wasted bandwidth and computational resources.

Web page access prediction gained its importance from the ever increasing number of e- commerce and e-business [Khalil et. al. 2007].The figure below shows web page prediction process.



It involves personalizing, Marketing, Recommendations, helps in improving the web site structure and also guide web users in navigating through hyperlinks for accessing the information they need. The most widely used techniques for discovering the patterns are Markov model, association rules and clustering,sequential patterns etc. However, each of the aforementioned techniques has its own limitations, especially when it comes to accuracy and space complexity [Khalil 2008].

In proposed work pre-fetching and prediction is done by pre-processing of logs as it is the main requirement to provide user with best recommendations[Cooley,Mobasher and Srivastava 1999] following three techniques together i.e. clustering, association rules and low- order Markov model using frequency support pruning, it achieves complete logs, better accuracy , less state space complexity and less number of rules. The predicted pages are pre- fetched and keep it in server cache which reduces the accessing time of that page and increases the web server performance. First of all a client request to a server for the specific web page. The server will send the URL of that page to the predictor. Then the predictor will check that specific web page, if it exists then predictor will send that page to the server and the server will immediately send that page to the client to fulfill its request. Also the predictor will send that page to the update engine which updates the data structure. The predictor uses that data structure for storing the web pages.

2.Literature Review

In 1998, R.Tewari, M. Dahlin, H. Vin and J. Kay examine several distributed caching strategies to improve the response time for accessing data over the Internet. By studying several Internet caches and workloads, Author derive four basic design principles for large scale distributed caches: (1) minimize the number of hops to locate and access data, (2) do not slow down misses, (3) share data among many caches, and (4) cache data close to clients.

In 1999, Reinhard P. Klemm, built a client-side Java-implemented prefetching agent, Web Companion, which employs a novel adaptive, fast, and selective online prefetching strategy based on estimated round-trip times for Web resources. This strategy efficiently hides the access latencies for slow resources while at the same time limiting the network and server overhead and local resource consumption to moderate levels.

In 2000, Greg Barish survey the state of the art in caching designs, presenting a taxonomy of architectures and describing a variety of specific trends and techniques.

In 2001, Jitian Xiao, presents an approach for measuring similarity of interests among Web users, based on the interest items collected from Web users access logs. A matrix-based algorithm is then developed to cluster Web users such that the users in the same cluster are closely related with respect to the similarity measure.

In 2001, Pablo Rodriguez discusses and compares the performance of different caching architectures. In particular, Author consider hierarchical and distributed caching. Author derive analytical models to study important performance parameters of hierarchical and distributed caching, i.e., client's perceived latency, bandwidth usage, load in the caches, and disk space usage.

In 2001, Jitian Xiao presents an approach for measuring similarity of interests among Web users from their past access behaviors. The similarity measures are based on the user sessions extracted from the user's access logs. A multi-level scheme for clustering a large number of Web users is proposed, the results obtained show that the clustering method is capable of clustering Web users with similar interests.

ISSN: 2278 – 1323

In 2003, Alexandros Nanopoulos present a new context for the interpretation of Web prefetching algorithms as Markov predictors. Author identifies the factors that affect the performance of Web prefetching algorithms. Author proposes a new algorithm called WMO, which is based on data mining and is proven to be a generalization of existing ones. It was designed to address their specific limitations and its characteristics include all the above factors. It compares favorably with previously proposed algorithms.

In 2005, Cairong Yan discusses that prefetching is an important technique for single Web server to reduce the average Web access latency and applying it on cluster server will produce better performance. Two models for parallel Web prefetching on cluster server described in the form of I/O automaton are proposed in this paper according to the different service approaches of Web cluster server: session persistence and session non-persistence.

In 2006, Junchang Ma gives a formal definition of the problem, presents an efficient and scalable algorithm for it. The algorithm has been implemented and applied to 16 large sets of Web pages. The experiments show that the algorithm can provide an average of 59.79%~72.28% bandwidth savings in fragment-based Web caching.

In 2007, Ruma Dutta presented an approach for storing Markov tree, used in various versions of PPM model while predicting next Web-page is proposed. Markov tree requires huge amount of memory. This problem is solved using Cellular Automata which is considered as a fast and inexpensive mechanism. The proposed technique utilizes non-linear Single Cycle Multiple Attractor Cellular Automata (SMACA) which replaces Markov tree for minimizing the memory requirement.

In 2008, Payal Gulati, proposes a Zipf's Law based novel approach for the determination of next page likely to be accessed by specific client.

In 2009, B. de la Ossa presents an empirical study to investigate the maximum benefits that Web users can expect from prefetching techniques in the current Web. To this end a perfect Web predictor is defined, but unlike previous theoretical studies, this work considers a realistic prefetching architecture using real and representative traces. In this way, the influence of real implementation constraints can be considered. The results obtained show that Web prefetching can improve page latency up to 52% in the studied traces.

In 2010, Yanjun Liu first describe an existing Web cache consistency technique which has been used in Web caching system. Later, Author proposes a strong cache consistency that is be suitable for Web appellations. Compared with previous strong consistency, Presented proposed algorithm is more efficient.

In 2011, Sina Bahram defined a work on the prediction of Web pages under the machine learning approaches. Author defined the structural and featured analysis on the Web pages to identify the individual and the relation features over the Web access. Author has defined three main datasets to perform the classification process. Author implemented the work in real environment and obtained results shows the effectiveness of the work [3].

In 2012, Hongyuan Ma performed a work on User-Aware Caching and Prefetching Query Results in Web Search Engines. In this paper author defined a User-Aware Cache, a novel approach tailored for query results caching that is based on user characteristics. Author also defined an approach for Web caching and then uses a trace of around 30 million queries to evaluate User-Aware Cache, as well as traditional methods and theoretical upper bounds. Experimental results show that this approach can achieve hit ratios better than state-of-the-art methods [9].

3.Proposed Model

3.1 Research Methodology

In this present work, an intelligent SOM improved neural network approach is defined to perform web usage mining and to predict the next visiting page. The work is about to improve the web access by performing the web prefetching. To perform pre-fetching, it is required to identify the current visiting page and predict the next visiting page. Once the page will be predicting, it can be cached before user demand. It will improve the web access as the next required page is already present in cache. To perform this web usage mining and web page prediction in this work a layered model is presented in this work. This layered model is shown here under

Layer 1 :

Clustering As the web access is performed in multiple sessions by number of users. There is a larger web log that stores the events performed by these users. This kind of web page visit is described in terms of time and visit number. In this work, at the initial stage, this web log will be analyzed under the frequency parameter. Because of which, the web pages will be categorized based on the usage. To perform this categorization, clustering will be performed on weblog dataset. From this dataset, high frequency cluster will be selected for further processing.

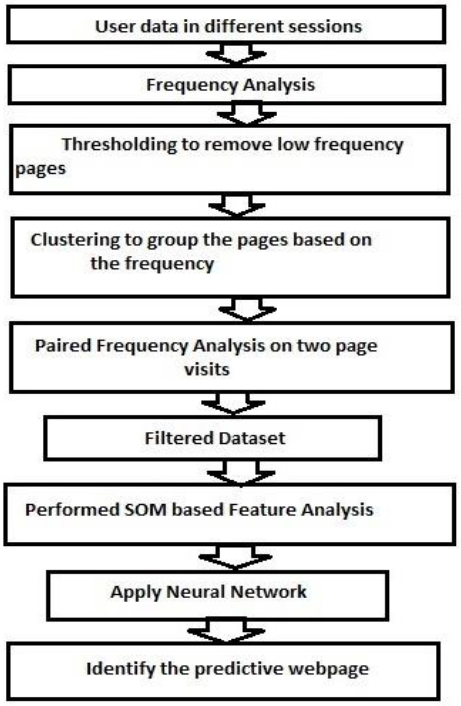
Layer 2 :

SOM In this stage, the analysis on the high frequency cluster will be performed using self organizing map. In this layer, the association between the visited webpages will be analyzed so that the linked visited patterns will be identified. This SOM will assign the weightage to various web pages based on predictive analysis.

ISSN: 2278 – 1323

Layer 3 :

Neural Network In this stage, the weighted web log contents will be trained under neural network to predict the next web page visit. The overall work flows through the steps given below



5.Result

Here figure 5.1 is showing the frequency measure obtained for different web pages. The SOM model is here applied to analyze the first page visit identification. Here x axis is showing the page number or the page itself and y axis is showing the possibility of page visit. The ratio based analysis is here performed to identify the page visit. The ratio is obtained respective to maximum visited page.

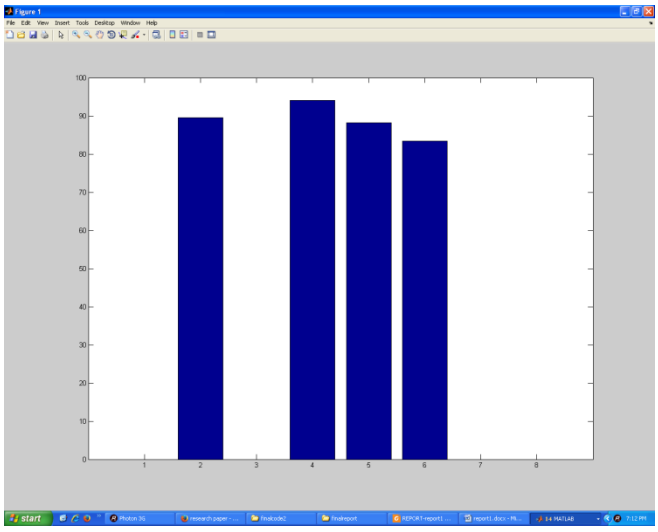


Figure 5.1 : Web Page Visit Probability (Page I)

Here figure 5.2 is showing the frequency measure obtained for different web pages for associated page 2. The SOM model is here applied to analyze the first page visit identification. Here x axis is showing the page number or the page itself and y axis is showing the possibility of page visit. The ratio based analysis is here performed to identify the page visit. The ratio is obtained respective to maximum visited page. As the first page is identified, the possibility of relative page visit is required to identify. The figure is showing the association adaptive analysis.

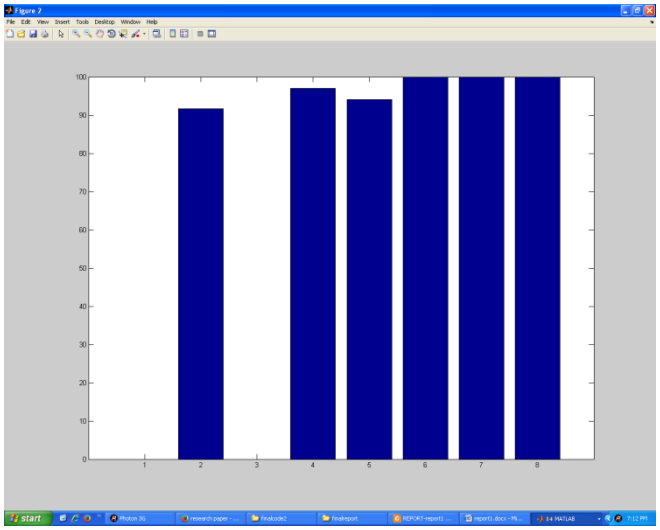


Figure 5.2 : Web Page Visit Probability (Assocaited Page)

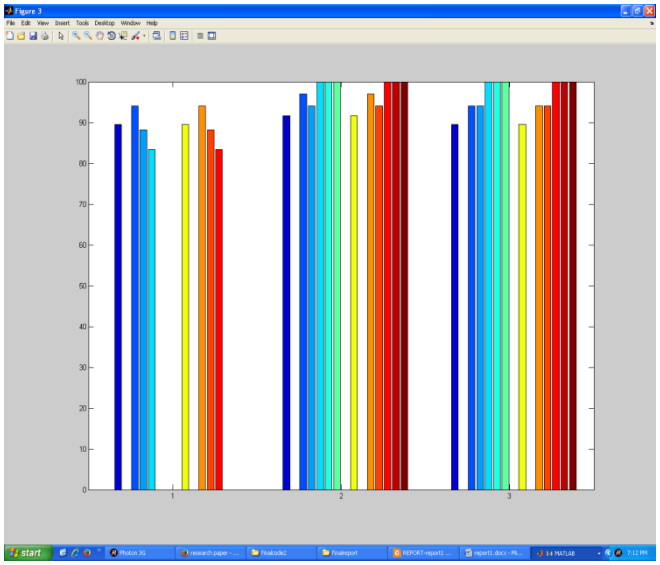


Figure 5.3 : Clustered Pages

Here figure 5.3 is showing the clustered results obtained from the work. The figure is showing the clustered results. Here 3 different clusters are formed. The figure is showing the cluster formation and the visit relative to the cluster. The figure shows that the cluster 2 and cluster 3 are having more frequent page visits.

ISSN: 2278 – 1323

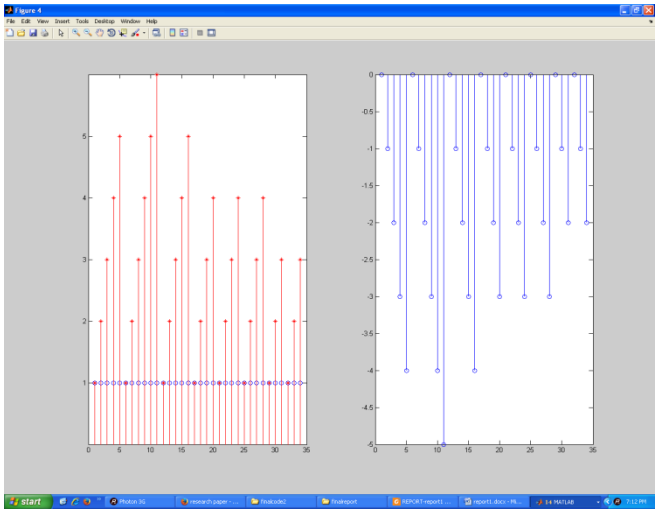


Figure 5.4 : Neural Predictive Page Results

Here figure 5.4 is showing the neural predictive results respective to different pages. Here figure is showing the page 10 is having the maximum frequency so that the chances of this page visit is high. After that the stage adaptive page visit is defined. Higher the line more frequent a page can be visited. The right side is the negative form of same model as the positive and negaitive weightages are processed.

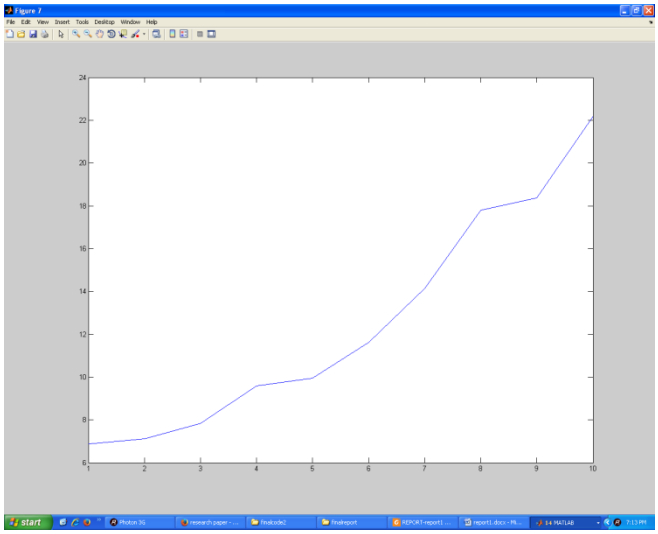


Figure 5.5 : Predictive Results

Here figure 5.5 is showing the predictive results obtanined in the form of possibility of next visiting page. The figure shows that the page 10 is having the highest probability for electing as the next page visit. So that the page can be considered to be cached.

6.Conclusion

Web page prediction is used by web browsers and many web servers to improve the web page access over the system. It can be implemented on client servers to improve the system performance. The web page prediction is about the analysis of user activity on the web server so that the next visiting page can be estimated before the user request and available it user cache so that the downloading and search time will be reduced. In this work, a SOM adaptive neural network model is defined to improve the system performance. The presented work is divided in three main stages. In first stage, the relation feature analysis is performed using SOM approach to identify the relation between the continous user requests. Once the requests are formed, the next stage is to categorize the visits under clustering approach, the fuzzy adaptive clustering is applied to generate the dataset segments. In final stage, the neural network model is applied to perform the next page prediction. The results shows that the work has provided the cluster view of next possible page visit.

7.References

[1] Yanjun Liu, “Strong Cache Consistency on World Wide We”, 2010 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE) [13] Venkata N. Padmanabhan, “Using Predictive Prefetching to Improve World Wide Web Latenc”, COMPUTER COMMUNICATION 1996

[2] Xiangping Chen,“ Lifetime Behavior and its Impact on Web Caching” Proceeding WIAPP IEEE, 1999.

[3] R. Tewari, M. Dahlin, H. Vin and J. Kay, “Beyond hierarchies: design considerations for distributed caching on the Internet", In Proceedings of the 19th International Conference on Distributed Computing Systems (ICDCS), 1998.

[4] Reinhard P. Klemm, “ WebCompanion: A Friendly Client-Side Web Prefetching Agent”, IEEE Transactions on knowledge and data engineering, 1999.

[5] Greg Barish and Katia Obraczka, "World Wide Web Caching: Trends and Techniques", IEEE Communications Magazine, May 2000.

[6] Jitian Xiao,” Measuring Similarity of Interests for Clustering Web-Users”, IEEE, 2001.

[7] Pablo Rodriguez, Christian Spanner, and Ernst W. Biersack, "Analysis of Web Caching Architectures: Hierarchical and Distributed Caching", IEEE/ACM Transaction on Networking, AUGUST 2001.

[8] Alexandros Nanopoulos, “A Data Mining Algorithm for Generalized Web Prefetching”, IEEE Transactions on knowledge and data engineering, 2003.

ISSN: 2278 – 1323

[9] Jitian Xiao, "Clustering of Web Users Using Session-based Similarity Measures", Proceedings. 2001 International Conference IEEE, 2001.

[10] Cairong Yan, "Parallel Web Prefectching on Cluster Server", CCECE/CCGEI, IEEE, 2005.

[11] M. Junchang, G. Zhimin, "Finding Shared Fragments in Large Collection of Web Pages for Fragment-based Web Caching", Fifth International Symposium on Network Computing and Applications (NCA'06), IEEE, 2006.

[12] Ruma Dutta, "Offering Memory Efficiency utilizing Cellular Automata for Markov Tree based Web-page Prediction Model", 10th International Conference on Information Technology IEEE, 2007

[13] Payal Gulati, "A Novel Approach for Determining Next Page Access", First International Conference on Emerging Trends in Engineering and Technology, IEEE, 2008.

[14] B. de la Ossa, "An Empirical Study on Maximum Latency Saving in Web Prefetching", International Conference on Web Intelligence and Intelligent Agent Technology, IEEE, 2009.

[15] Yanjun Liu, "Strong Cache Consistency on World Wide Web", 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), 2010.

[16] Sina Bahram, "Prediction of Web Page Accessibility Based on Structural and Textual Features", Co-Located with the 20th International World Wide Web Conference, W4A2011 - Communications paper, March 28-29, 2011.

[16] Xiangping Chen," Lifetime Behavior and its Impact on Web Caching" Proceeding WIAPP IEEE, 1999..

[17] Hongyuan Ma," User-Aware Caching and Prefetching Query Results in Web Search Engines", SIGIR'12, August 12–16, 2012.

ISSN: 2278 – 1323