Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 5, May 2015, pg.175 – 181

RESEARCH ARTICLE

SOM Improved Neural Network Approach for Next Page Prediction

Vidushi
Research Scholar, Department of Computer Science and Engineering, Ganga Institute of Technology & Management, Kablana
Email ID: gulia.vidushi@gmail.com

Dr. Yashpal Singh
Asst. Professor (CSE Dept.), Ganga Institute Of Technology & Management, Kablana
Email ID: yashpalsingh009@gmail.com

ABSTRACT-

The increasing usage of web results the heavy communication and slow returns from web. Because of this, there is the requirement of some approaches to optimize the web resources usage. One of such approach is caching that can be used within an organization to optimize the access of frequently used web pages. Caching is about to predict the requirement of next web access of a user and load it in cache before user request. This kind of intelligent prediction comes under web usage mining. In this work, an intelligent SOM(Self Organizing Map) improved neural network approach is defined to perform next web page prediction. The work will be here presented in three main stages. In first stage, to perform the intelligent sequence mining the dataset will be filtered. The filtration will be here performed using clustering approach. The clustering will be performed based on web usage. Now only the cluster that represents the high usages pages will be considered for prediction. In second stage, SOM will be applied to analyze the web page usage and the prediction of next required page access. The SOM will be applied here to assign the weightage to next possible based on frequency and time stamp analysis. Once the weightage will be applied, the final work is to apply the neural network to predict the next visiting page.

GENERAL TERMS- Web Caching, Web prefetching, web usage mining

KEYWORDS- SOM, WWW, HTML

1. INTRODUCTION-

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.[12]

The rapid growth of the WWW inspired numerous techniques to reduce web latency. The different techniques in latency reduction are:

1. Web Caching

2. Pre-fetching

3. Preopening

Web caching is recognized as one of the effective techniques to alleviate the server bottleneck and reduce network traffic, thereby reducing network latency. The basic idea is to cache recent requested pages at the server so that they do not have to be fetched again. Regular caching however, only deals with previously requested files, i.e. by definition, new files will never be in the cache[12].

Web Pre-fetching, which can be considered as "active" caching, builds on regular Web caching and helps to overcome its inherent limitation. It attempts to guess what the next requested page will be. For regular HTML file accesses, pre-fetching techniques try to predict the next set of files/pages that will be requested, and use this information to pre-fetch the files/pages into the server cache. This greatly speeds up access to those files, and improves the users' experience. To be effective however, the pre-fetching techniques must be able to reasonably predict (with minimum computational overheads) subsequent web accesses[12].

Web Preopening is how to speculatively pre-open network connections. Instead of pre- fetching all files, client only makes the DNS lookup and sets up the three-way handshake with server. This can reduce latency to some extent and have a small risk[12].

WEB CACHING The long-term success of the World Wide Web depends on fast response time. People use the Web to access information from remote sites, but do not like to wait long for their results. The rapid growth in the amount of information and the number of users has lead to difficulty in providing effective response time for the web users and this increased web latency; resulting in decreased web performance. Although several proposals have been made for reducing this latency, like it can be improved by caching, the benefit of using it is rather limited owing to filling the cache with documents without any prior knowledge. Predictive caching becomes an attractive solution wherein the forthcoming page likely to be requested soon are predicted based on user access logs information and pre-fetched ,while the user is browsing the current display pages. As web page prediction gained its importance, therefore in this thesis we develop a novel technique to implement web page prediction process by pre- processing the user access log and integrating the three

techniques i.e. Clustering, Markov model and association rules which achieves better web page access prediction[13].

DEFINITION Web caching is the storage of Web objects near the user to allow fast access, thus improving the user experience of the Web surfer. Examples of some Web objects are Web pages (the HTML itself), images in Web pages, etc. [Davison 2001].

2. PREDICTIVE CACHING/PREFETCHING

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]. 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] and also overcomes the limitation of path completion and for pattern discover we integrate the 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.

3. LITERATURE REVIEW-

3.1 STUDY OF USE OF PATTERN DISCOVERY TECHNIQUES FOR WEB PAGE PREDICTION

Web page access prediction gained its importance from the ever increasing number of e- commerce and e-businesses[Khalil et. al. 2007]. It involves personalizing, Marketing, Recomendations, helps in improveing 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. we highlight here the

significance of studying the pattern discovery techniques for evolving nature of the Web page prediction.

Q. Yang [Yang et al 2004] studied different association rule based methods for web request prediction. Using association rules for web access prediction involves dealing with too many rules and it is not easy to find a suitable subset of rules to make accurate and reliable predictions.He has studied five different representations of Association rules which are: Subset rules, Subsequence rules, Latest subsequence rules, Substring rules and Latest substring rules. As a result of the experiments, performed by the authors concerning 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.

B. Liu[Liu et al 1998] [Yang et al 2004] have introduced a customized marketing on the Web approach using a combination of clustering and association rules. The authors collected information about customers using forms, Web server log files and cookies. They 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. They then performed association rule techniques on each cluster.

D. Kim[Kim et al. 2004][Chen et. al.2008] combined all three models together . It improve 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.His work improved recall and it did not improve the Web page prediction accuracy.

Vakali et al. [4] categorize 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 [5]. 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.

3.2 STUDY OF INTELLIGENT TECHNIQUES OF WEB PAGE PREDICTION

This section correspond to the study of various techniques which had been deployed for web page prediction.

Zukerman et al in [6], uses Artificial Intelligence-related techniques to predict user requests. They 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.

Victor Y. Safronov presented the Page Rank based prefetching 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 prefetching. 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.

Padmanabhan et al [7] investigated ways of optimizing retrieval latency. Web caching has been recognized as an effective solution to minimize user access latency A method of called prefetching was introduced in which clients in collaboration with servers prefetch web page that the user is likely to access soon, while he/she is viewing the currently displayed pagel [8]. The benefit of prefetching is to provide low retrieval latency for users, which can be explained as high hit ratio Investigate an approach to reduce web latency by prefetching between caching, proxies, and browsers. Research on predictive Web prefetching has involved the important issue of log file processing and the determination of user transactions (sessions) from it [6,9 ,10,11 ,12] provide various data mining algorithms for the path traversal patterns and how to efficiently mine the access patterns from the web logs.

3.3 STUDY OF PREFETCHING STRATEGIES

This section correspond to the study of various strategies involved till now for prefetching web pages.

Jose Borges and Mark Levene[Borges and Levene 2004] propose a dynamic clustering-based method to increase a Markov models accuracy in representing a collection of user web navigation sessions. The method makes use of the state cloning concept to duplicate states in a way that separates in-links whose corresponding second-order probabilities diverge. In addition, the new method incorporates a clustering technique which determines anecient way to assign in-links with similar second-order probabilities to the same clone.

Siriporn Chimphlee [Chimphlee et. al. 2006] present a rough set clustering to cluster web transactions from web access logs and using Markov model for next access prediction. Using this approach, users can effectively mine web log records to discover and predict access patterns. He perform experiments using real web trace logs collected from www.dusit.ac.th servers. In order to improve its prediction ration, the model includes a rough sets scheme in which search similarity measure to compute the similarity between two sequences using upper approximation.

Silky Makker and R.K Rathy[Makker and Rathy 2011] proposes a bracing approach for increasing web server performance by analyzing user behavior, in this pre-fetching and prediction is done by pre-processing the user access log and integrating the three techniques i.e. Clustering, Markov model and association rules which achieves better web page access prediction accuracy;This work also overcomes the limitation of path completion i.e. by extracting web site structure paths are completed,which helps in better prediction, decreasing access time of user and improving web performance.

3.4 STUDY OF DIFFERENT MODELS USED FOR PREFETCHING

This section correspond to the study of various algorithm which has been used in the various stages of web page prediction process i.e.

R.R. Sarukkai [Sarukkai 2000] used first-order Markov models to model the sequence of pages requested by a user for predicting the next page accessed. A "personalized" Markov model is trained for each individual and used for predictions in user's future request sessions. In practice, however, it is very expensive to construct a unique model for each user respectively, and the problem gets even worse when there exist thousands of different users within a big Web site.

V. Padbanabham and J. Mogul [Padbanabham and Mogul 1996] use N-hop Markov models predicted the next web page users will most likely access by Pmatching the user's current access sequence with the user's historical web access sequences for improving prefetching strategies for web caches.

4. CONCLUSION-

- The session based clustering process will reduce the analysis dataset so that the efficiency of system will be improved.
- The neural network based approach will be able to provide the more intelligent and accurate prediction.

REFERENCES

[1] Bhawna Nigamand Dr. Suresh Jain, "ANALYSIS OF MARKOV MODEL ON DIFFERENT WEB PREFETCHING AND CACHING SCHEMES", 978-1- 4244-5967-4/10/ ©2010 IEEE

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

[3] S. Yang, J. Zhang and S. Tsai, "An automatic Semantic- Segment Detection Method in the HTML Language", Services Computing, 2008. SCC apos;08. IEEE International Conference on Volume 1, Issue , 7-11 July 2008.

[4] S. Yu, D. Cai, J.-R. Wen, and W.-Y. Ma, "Improving pseudo- relevance feedback in web information retrieval using web page segmentation", In Proceedings of the Twelfth International World Wide Web Conference, WWW2003, pp. 11-18, Budapest, Hungary, May 20-24, 2003.

[5] Jia Wang, "A Survey of Web Caching Schemes for the Internet" ACM SIGCOMM 2000.

[6] P. Cao, J. Zhang, and K. Beach, "Active Cache: Caching dynamic contents on the web"

[7] R. Tewari, M. Dahlin, H. Vin and J. Kay, "Beyond hierarchies: design considerations for distributed caching on the Internet"

[8] Brian D.Davison, "A Web Caching Primer" IEEE INTERNET COMPUTING 2001

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

[10] Pablo Rodriguez, Christian Spanner, and Ernst W. Biersack, "Analysis of Web Caching Architectures: Hierarchical and Distributed Caching", IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 9, NO. 4, AUGUST 2001

[11] L. Ramaswamy, A. lyengar, L. Liu, F. Douglis, “Automatic Fragment Detection in Dynamic Web Pages and Its Impact on Caching”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 17, NO. 6, JUNE 2005.

[12] 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

