International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Hybrid Prefetching for Improving Page Hit-Ratio

Kushwant Kaur[1], Dr. Kanwalvir Singh Dhindsa[2]

[1]Research Scholar, Department of CSE & IT, BBSBEC,Fatehgarh Sahib, Punjab, India

[2]Associate Professor, Department of CSE, BBSBEC, Fatehgarh Sahib, Punjab, India

Abstract: ***Web prefetching techniques help in improving web page hit-ratio and reducing user perceived latencies by fetching in advance, the most useful web objects into proxy sever cache. This paper explains the integrated prefetching approach that combines Domain top prefetching and Dynamic web prefetching. The data obtained from this approach is based onuser access patterns and is stored into the proxy server cache to solve the problem of cache replacement and as a result the hit-ratio is increased.***

Keywords: Web prefetching, user-perceived latency, Domain-Top Prefetching, Dynamic Web prefetching.

1. Introduction

The rapid growth in use of Internet and World Wide Web leads to decrease in latency rates. Web Caching improves the quality of Web service [21] by storing the web objects closer to the user for faster access. But the cache replacement problem lies in the heart of Web Caching, which describes that a new page will replace the existing page in cache and hence results in a miss if the demanded page is not found in the cache. Many cache replacement algorithms have been proposed for ranking web objects based on their access patterns.

In today's world, the visitor is least patient as he can't wait for the server to respond to his request and he shifts to another website if the requested one is slow responding. So, it is a challenging task to speedup the web page response. This can be achieved with the development of a predictive model, which can be used to forecast web objects that are most likely to be visited by the user in future. This predictive technique is known as Web Prefetching

2. Web Pre-Fetching

Web Prefetching means retrieving data from remote servers in expectation of client requests.Web prefetching can be applied in three ways; 1) between clients and web server, 2) between proxy servers and web server and 3) between clients and proxy server [11]. The problem of user perceived latency can be solved by applying prefetching between clients and web serverbut it will increase the network traffic. It predicts the web objects most likely to be accessed by the user and fetches them into cache. The web prefetching predicts the web objects that are expected to be requested by the user in the near future. This is based on the usage patterns of the visitors. This technique fetches the urls of the websites from the server and based on their demand these can be clustered in a group of similar type users. Many web-prefetching techniques are available to perform the task of advance fetching.

2.1 Domain Top Pre-fetching

Domain top approach for web prefetching associates the proxy's active knowledge of most popular domains and documents[11],[13]. In this, the proxy calculates the most popular domains and most popular documents in those domains along with rank list for prefetching. The most frequently visited web objects are placed at the top and least recently visited is placed at the bottom.

2.2 Dynamic Web Pre-fetching

Dynamic web pre-fetching technique stores a preference list in proxy server's database[12]. Intelligent agents are used for parsing the web page, monitoring the bandwidth usage and maintaining hash table, preference list and cache consistency. It controls the web traffic by reducing pre-fetching at heavy traffic and increasing pre-fetching at light traffic. Thus it reduces the idle time of the existing network and makes the traffic almost constant. A hash table is maintained for storing the list of accessed URLs and its weight information [22]. Depending upon the bandwidth usage and weights in the hash table, the prediction engine decides the number of URLs to be pre-fetched and gives the list to pre-fetch engine for pre-fetching the predicted web pages. After pre-fetching, the proxy server keeps the pre-fetched web pages in a separate area called pre-fetch area.

2.3 Integration of Domain Top and Dynamic Web Prefetching

Dynamic web pre-fetching and Domain Top approach of pre-fetching can be combined to improve web page response time. Hence, preference list from Dynamic technique can be brought into Domain Top approach. Optimized top domain approach consists of preference list along with the rank list. Advantage of using approach this is that the prefetching would have wider scope as being be more fast. Due to presence of rank list (which maintain top domain and top documents) and preference list (which consist of list of sites stored by user to have immediate access) model will work well. This model covers the loopholes like overhead in proxy servers.

3. Implementation Model

The implementation model for proposed hybrid technique has focused on improving the web page response time and increasing page hit-ratio by enhancing web pre-fetching process. For experimentation we have used database with various web entries and have done cleaning process on the

database. We have used the web log file having 65,536 entries. The url entry in the log file is in the form of a string like http://www.facebook.com/friends/ajax/lists.php.Data cleaning removes the php link and additional user details. In data cleaning all the images, jsp pages and the user shown data is just terminated. After cleaning we are left with www.facebook.com.Apriori algorithm has been used to display the user information. The result is a list of urls corresponding to domains along with their rank list.By implementing the hybrid technique of web preprocessing the following sample data given below has been extracted with the rank list:

Top 10 list of the edu website
http://imke2012.chitkara.edu :::::: 40
http://its.edu :::::: 40
http://imke2012.chitkara.edu :::::: 40
http://its.edu :::::: 40
http://images.sbbs.edu :::::: 30
http://www.isi.edu :::::: 15
http://www.grad.buffalo.edu :::::: 12
http://www.chem.uwec.edu :::::: 9
http://www.chitkara.edu :::::: 8
http://phobos.ramapo.edu :::::: 4

Top 10 list of the com website
http://www.facebook.com :::::: 3384
http://ad.yieldmanager.com :::::: 1015
http://l.yimg.com :::::: 936
http://www.google-analytics.com :::::: 910
http://www.bhaskar.com :::::: 724
http://armdl.adobe.com :::::: 722
http://safebrowsing-cache.google.com :::::: 685
http://www.google.com :::::: 579
http://www.scopus.com :::::: 563
http://cdn.fetise.com :::::: 543

Top 10 list of the ac.in website
http://campus.pu.ac.in :::::: 752
http://www.ignou.ac.in :::::: 198
http://puchd.ac.in :::::: 155
http://mail6.pu.ac.in :::::: 97
http://pumail.puchd.ac.in :::::: 65
http://uiet.puchd.ac.in :::::: 45
http://jobs.puchd.ac.in :::::: 21
http://www.uiet.puchd.ac.in :::::: 17
http://news.puchd.ac.in :::::: 15
http://pumail.pu.ac.in :::::: 14

3.1 Hybrid Web Prefetching Algorithm

The following algorithm has been used as preprocessing hybrid technique

1. Start fetching the URLs the user needs to visit.
2. In the backend the information of user as per his IP address will be stored in a SQL server database file that contain IP address, URL and time.
3. Now the data of the first user is combined he creates a preference list on the basis of visit to a site.
4. On the other hand a domain top list is created which contains the top domains with their respective top documents with rank.
5. Now when the user will enter the web second time, a list of URLs, previously fetched by him on the basis of their rank in Domain list, will be shown to him that he can directly fetch.
6. If the user found the requested document in the list shown then it is known as a hit else a miss and request is forwarded to origin server.

In this module we have considered three users named user1, user2, user3 at the same level. The cache size is 25 for the users, which means that if the size of the user requirement increases more than 25 then only most frequent pages will be displayed in the buffer.

Example: If the user1 is having 19 pages those pages will be displayed in the cache. Now if the user 2 will request for the page and he is having 14 pages his only 6 pages will be brought in the cache. If now again user1 request for the pages then in this case the pages will be repaved with the most frequent pages from the database that is stored with the timestamp in the data base.

The graph for the user1, user2, user3 is based on the hit ratio that is the number of timesthe user entered in the website to the number of time the match found in server.

4. Results and Discussions

The results are analyzed on the basis of number of times the request for web pages made and the number of times the requested page is found in the preference list. If the page is found in the list then it is considered as a hit ratio. We have experimented the hybrid technique on our web log file of sample data and obtained the following results represented in Figure 1, Figure 2 and Figure 3.

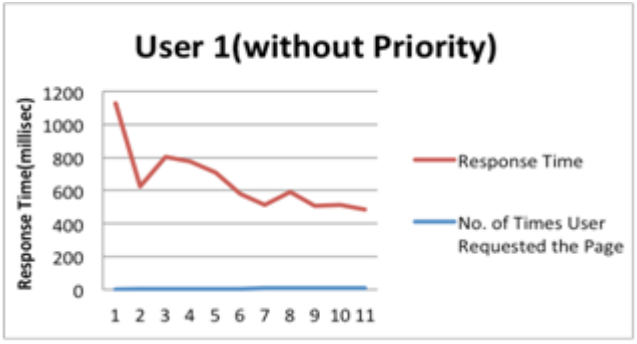

Figure 1: Hit-ratio for user1

The resultant graph shown in figure 1 represents hit-ratio of user 1 i.e. the number of timesthe user1 entered in the website to the number of time the match found in server response time.When the user1 requested an already visited website for the 2nd time then the server responded in 600 milliseconds and when he requested for the same website for the 10th time then the server responded within only 450 milliseconds. It shows that as the user is requesting a given site multiple times then that website is kept in the buffer/cache for a longer time because it has been pre-fetched on the basis of user1 demands. Similarly, for those websites, which have been previously visited by the user multiple times, the response time of the server will be higher.

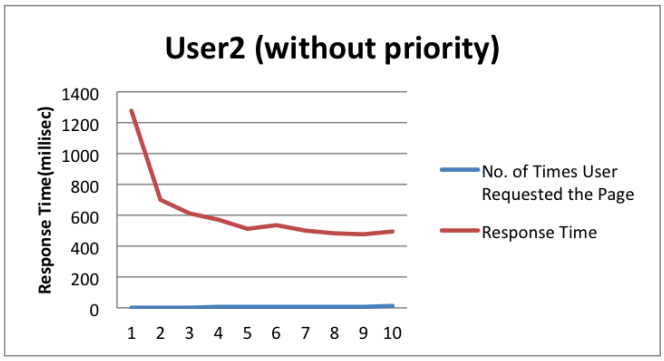

Figure 2: Hit-ratio for user2

The resultant graph shown in figure 2 represents hit-ratio of user2 i.e. the number of timesthe user2 entered in the website to the number of time the match found in server response time.

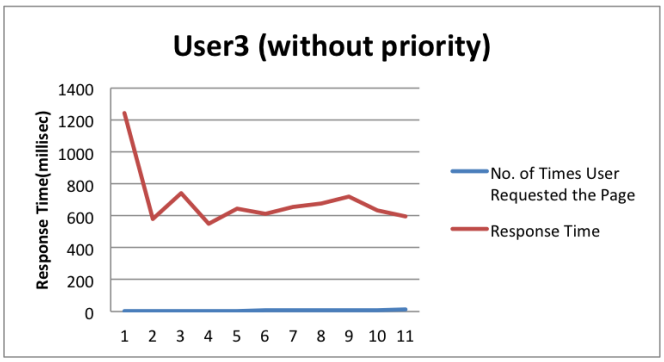

Figure 3: Hit-ratio for user3

Similarly, the resultant graph shown in figure 3 represents hit-ratio of user3 i.e. the number of timesthe user3 entered in the website to the number of time the match found in server.

Web Prefetching can be implemented by giving priority to specific user so that the users with crucial task are addressed with highest speed.

5. Conclusions

In this paper, we have described integration of Domain Top Web Prefetching and Dynamic Web Prefetching for improvement of web page response time. By storing the most frequently requested future web objects closer to the users, cache miss ratio can be dropped and hit-ratio can be increased. Prefetching can be implemented either by giving priority to a specific user or by treating all users equally based on the application area.

References

[1] I. Dzitac, "Advanced AI techniques for web mining," Proceedings of the 10th WSEAS international conference on Mathematical methods, computational techniques and intelligent systems.Corfu Greece. 2008.
[2] J. B. Patil and B. V. Pawar, "Improving Performance on WWW using Intelligent Predictive Caching for Web Proxy Servers," IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 1, January 2011.
[3] AlexandrosNanopoulos, DimitriosKatsaros, and YannisManolopoulos, "Exploiting Web Log Mining for Web Cache Enhancement," LNAI 2356, Springer-Verlag Berlin Heidelberg , pp. 68–87, 2002.
[4] Siddharth Jain, Ruchi Dave, Devendra Kumar Sharma, " An approach using Association Rule Mining Technique for frequently matched pattern of an Organization's web log data", International Journal of Engineering Sciences & Research Technology, Vol.1, No.5, pp 297-300, 2012.
[5] HendrikBlockeel, Raymond Kosala," Web Mining Research: A Survey", ACM SIGKDD, Vol- 2, Issue- 1, pp 1, 2000.
[6] P. Somrutai, "Improving the Performance of a Proxy Server using Web log mining," M.S. thesis, San Jose State University, 2011.
[7] RudeekornSoonthornsutee, PramoteLuenam," Web Log Mining for Improvement of Caching Performance ", Proceedings of the International Multi-conference of Engineers and Computer Scientists, Vol - 1, pp 14-16, March 2012.
[8] E AlexandrosNanopoulos, DimitriosKatsaros, and YannisManolopoulos, "Exploiting Web Log Mining for Web Cache Enhancement", WEBKDD 2001, LNAI 2356, pp. 68–87, 2002.
[9] J. Srivastava, R. Cooley, M. Deshpande, P.-N. Tan," Web usage mining: discovery and applications of usage patterns from web data", SIGKDD Explorations, Vol – 1, Issue- 2, pp 12–23, 2000.
[10] JosepDome`nech, Ana Pont, Julio Sahuquillo, Jose´ A. Gil," A user-focused evaluation of web prefetching algorithms", Computer Communications, ScienceDirect, Vol- 30, pp 2213-2224, 2007.
[11] Seung Won Shin, Byeong Hag Seong&Daeyeon Park, (2000)"Improving World-Wide- Web Performance Using Domain- Top Approach to Prefetching", Fourth International Conference on High-Performance Computing in the Asia-Pacific Region vol. 2, pp. 738-746.
[12] V. Sathiyamoorthi, V. MuraliBhaskaran," Improving the Performance of Web Page Retrieval through Pre-Fetching and Caching using Web Log Mining ", European Journal of Scientific Research, Vol.66, No.2, pp. 207-218, 2011.
[13] Greeshma G. Vijayan and Jayasudha," A Survey on Web Pre- fetching and Web Caching Techniques in a Mobile Environment", ITCS, SIP, JSE-2012, CS & IT 04, pp. 119–136, 2012
[14] Yin-Fu Huang, Jhao-Min Hsu," Mining web logs to improve hit ratios of prefetching and caching", Knowledge- Based Systems, Science Direct, Vol- 21, pp 62-69, 2008.
[15] Walees Ali, SitiMariyanShamsuddin and Abdul Samad Ismail, "A Survey of Web caching and Prefetching", Int. J, Advance Soft Comput. Appl., Vol. 3 No.1, March 2011.
[16] U. Acharjee, "Personalized and Artificial Intelligence Web Caching and Prefetching",Master Thesis, University Ottawa, Canada.
[17] P.Somurthai, "Improving the Performance of a Proxy Server using Web log mining" M.S. thesis, San Jose State University.
[18] V. Sathiyamoorthi and Dr.Murali, " Data Preprocessing Techniques for Pre-Fetching and Caching of Web Data Through Proxy Server", International Journal Of

Computer Science and Network Security, Vol 11, No.11, 2011.
[19] George Pasllis , Athena Vakali, PokornyJaroslav, "A clustering- based prefetching scheme on a Web cache environment', www.sciencedirect.com Computers and Electrical Engineering 34,pp. 309-323.
[20] Suresha, " Caching Techniques For Dynamic Web Servers", Ph. D. Thesis, Indian institute Of science, Banglore, 2007.
[21] Q. Yang, H. Zhang "Integrating Web prefetching and caching using prediction models", World Wide Web, pp. 299–321,2001.