ISSN:0975-9646

Hybrid Approach for Improvement of Web page Response Time

Kushwant Kaur[#], Prof. Kanwalvir Singh Dhindsa[*]

[#]*Research scholar, Dept. Of CSE,*
BBSBEC(Fatehgarh Sahib), Punjab(India)

[*]*Associate Professor, Dept Of CSE*
BBSBEC (Fatehgarh Sahib), Punjab (India)

***Abstract*— The objective of this paper is to propose an integrated web mining technique for improving the response time of web pages and reducing user perceived latency. This hybrid technique is based on integration of two Web mining techniques: Web caching and Web pre-fetching. By integration, these techniques may compliment each other since Web Caching exploits temporal locality and Web Pre-fetching utilizes spatial locality. Pre-fetching the predicted web objects into the proxy server cache can increase cache hit-ratio. A comparison between web pre-fetching techniques is done to present the benefits of hybrid approach. This paper also proposes that if Prefetching techniques like Domain Top Prefetching and Dynamic Web Prefetching are combined then it can further improve the Hybrid approach.**

***Keywords*— Web mining; user perceived latency; proxy server; Web caching; Web pre-fetching; temporal locality; spatial locality; cache hit-ratio; Domain Top prefetching; dynamic prefetching.**

I. INTRODUCTION

This Web has become today a virtual society as being used for communication channels and search purposes. Network congestion has become the biggest problem due to the massive use of Internet and World Wide Web. Server is overloaded with user's frequent requests for web page access. This dramatic increase is due to popularization of new applications and services like e-commerce, e-learning, e-business, multimedia contents etc. and has given rise to problems like user perceived latency, global traffic, damaging the quality of service and backbone link congestion. User perceived latency has been considered as the most serious problem as it results in impatience which is the most common reason users terminate their visit at web sites. Potential sources of latency are the overloaded web servers, network congestion, low bandwidth, bandwidth underutilization and propagation delay.

Web mining can be used to improve latency. Web mining is a type of data mining used to automatically discover and extract information from Web documents and other web services. It consists of specific techniques, algorithms and methodologies to mine the web, mainly because the web has a great amount of unstructured data and the changes are frequent and rapid. Web mining techniques can be implemented on the web logs maintained byte servers as to discover user access and traversal paths.

Web caching has been used as one of the effective techniques to reduce network traffic, access latencies and bandwidth underutilization. [4] Cache storage space is limited and some pages need to remove when cache is full and the new pages are to be brought into the cache. This cache replacement may lead to inefficiency as the deleted page may be requested again. A lot of studies have been done o improve Web caching performance. For example, [6] applied a technique of predicting the future web access using Web Log Mining for improvement of caching performance. An integrated techniques has been suggested in [8] as an approach of clustering and classification techniques for the building the model of cache replacement policy.

The knowledge and comprehension of the behavior of a web user are important factors in a wider range of fields related to web architecture, design, and engineering. The information that can be extracted from web user's behavior permits to infer and predict future accesses. This information can be used for improving Web usability as proposed by [3] developing on-line marketing techniques suggested by [9] or reducing user-perceived latency, which is the main goal of prefetching techniques. These techniques use access predictors to process a user request before the user actually makes it.

The goal of the paper is to increase the hit ratio by proxy pre-fetching and lessen the burden on the proxy server and the network. In our scheme, we propose an integrated approach for improvement of web performance and to reduce the web page response/retrieval time. In Domain Top technique of pre-fetching; the proxy finds the popular domains using access profiles and searches the popular documents in each domain [11]. Based on these Top-Domain and Top-Documents, proxy makes the rank list for prefetching, the client requests a file in a certain domain and proxy forwards to them their most popular documents in the rank list. Heavy computation is not required to find the popular domains and documents, but only needs a very small amount of rank list that stores them at the proxy.

The paper is organized as follows. A brief description of Web log mining process is reviewed with the help of sample data taken from IIS and apache server logs in Section 2. Section 3 describes the web mining techniques for performance improvement along with the integration of

Web Caching and Web Prefetching techniques. It also discusses the results of comparative study of Web Mining Techniques. The schematic architecture is proposed for implementation of improved hybrid approach that consists of a combination of Domain top and Dynamic Pre-fetching in section 4. Finally, the conclusion in stated in Section 5.

II. AN OVERVIEW OF WEB LOG MINING

An 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 web users by reorganization of website. The mined data consists of the identity or origin of Web users data logs of user's web interaction, web server logs, proxy server logs and browser logs, data about referring page, user spent time at site and sequence of pages visited. This is important to the overall use of web mining for companies and their internet/ intranet based applications and information access.

Web logs have been considered as the best tools by [17] for understanding customers as these helps on knowing how the customers find your website and why they are looking for it. The mind-set of arriving visitors is best judged with this information. We can analyse web logs to know the access patterns of users, which provides information regarding most frequently visited web pages of our website. Clustering is performed on the similar type of access patterns. These groups help in determining pages liked by users whenever he logs into the site. Web pre-fetching technique fetches these most likely future pages into the proxy server cache and hence increases the access speed and reduces user perceived latency.

The Web log file is text file, which stores records in identical format. Each record in the log file represents a single HTTP request. A log file record contains important information about a request: the client side host name or IP address, the date and time of the request, the requested file name, the HTTP response status and size, the referring URL, and the browser information. Each Web server has its own log file format. Internet Service Provider (ISP) hosted files may not keep the log files for you, because log files can be very huge if the site is very busy. Instead, they only give you statistics reports generated from the logs files.
Here are some sample records from an IIS server log file:
02:49:12 127.0.0.1 GET / 200
02:49:35 127.0.0.1 GET /index.html 200
03:01:06 127.0.0.1 GET /images/sponsered.gif 304
03:52:36 127.0.0.1 GET /search.php 200
04:17:03 127.0.0.1 GET /admin/style.css 200
05:04:54 127.0.0.1 GET /favicon.ico 404
05:38:07 127.0.0.1 GET /js/ads.js 200

The simple format has fields for: time, client IP address, request command, requested file, and response status code. The format for apache server is more complex as compared to IIS. Some samples taken from Apache in its raw format are shown below:

192.168.198.92 - - [22/Dec/2002:23:08:37 -0400] "GET / HTTP/1.1" 200 6394 www.yahoo.com "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1...)" "-"

192.168.198.92 - - [22/Dec/2002:23:08:38 -0400] "GET /images/logo.gif HTTP/1.1" 200 807 www.yahoo.com "http://www.some.com/""Mozilla/4.0 (compatible; MSIE 6...)" "-"

192.168.198.92 - - [22/Dec/2002:23:08:40 -0400] "GET /images/wall.jpg HTTP/1.1" 200 807 www.yahoo.com "http://www.some.com/""Mozilla/4.0 (compatible; MSIE 6...)" "-"

192.168.72.177 - - [22/Dec/2002:23:32:14 -0400] "GET /news/sports.html HTTP/1.1" 200 3500 www.yahoo.com "http://www.some.com/""Mozilla/4.0 (compatible; MSIE...)" "-"

192.168.72.177 - - [22/Dec/2002:23:32:14 -0400] "GET /favicon.ico HTTP/1.1" 404 1997 www.yahoo.com "-" "Mozilla/5.0 (Windows; U; Windows NT 5.1; rv: 1.7.3)..." "-"

192.168.72.177 - - [22/Dec/2002:23:32:15 -0400] "GET /style.css HTTP/1.1" 200 4138 www.yahoo.com "http://www.yahoo.com/index.html" "Mozilla/5.0 (Windows..." "-"

192.168.72.177 - - [22/Dec/2002:23:32:16 -0400] "GET /js/ads.js HTTP/1.1" 200 10229 www.yahoo.com"http://www.search.com/index.html" "Mozilla/5.0 (Windows..." "-"

192.168.72.177 - - [22/Dec/2002:23:32:19 -0400] "GET /search.php HTTP/1.1" 400 1997 www.yahoo.com "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; ...)" "-"

These samples provide information that can be traced from such logs, and to a limited extent how this could impact on privacy when surfing.

Here is a sample line taken from the above sample web log in its raw format:

192.168.72.177 - - [22/Dec/2002:23:32:19 -0400] "GET /search.php HTTP/1.1" 400 1997 www.yahoo.com "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1;...)" "-"
This web server log line tells us:

- Visitor's IP address or hostname [192.168.72.177]
- Login [-]
- Authuser [-]
- Date and Time [22/Dec/2002:23:32:19 -0400]
- Request Method [Get]
- Request Path [search.php]
- Request protocol [HTTP/1.1]
- Response Status [400]
- Response content size [1997]
- Referrer path [www.yahoo.com]
- User agent [Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1;...)]

Considerations made from the web log file:

- Each line in the file represents a single "hit" on the web server which is different from a web page hit because a web a visit to web page may consist of various web objects like jpg images, gif images and other content etc. For example the visit by "192.168.198.92" consists of 1 gif and 1 jpg image.
- The entries of log are not in strict chronological order. Log may be updated after completion of every transfer. For example the main page completes after images on that page.
- One site may consist call to another site for retrieval of web content.
- In real applications the visitors may overlap which is not shown in the log file.

[9] Proposed a technique of clustering users based on their web access patterns.

III. Web Mining Techniques for Improvement of Web Performance

Web caching and Web mining have been considered as most popular techniques of web mining that play a vital role in improving the web performance by keeping the most frequently visited web objects closer to the client.

A. Web Caching

Web Caching is a web mining technique used for the improvement of performance of web-based systems. Performance is improved by storing the most likely visited web objects closer to the client machine or proxy server. Web Caching offers following advantages [16]:

- Decreases user perceived latency.
- Reduces network bandwidth usage.
- Reduces loads on the origin servers.

The combination of these features make caching technology attractive to all Web participants, including end-users, network managers, and content creators. Web caching keeps a local copy of web pages. Caches are found in browsers as well as in intermediary positions between the user agent and the origin server. Whenever user requests a web page, its availability is first checked in cached pages. If web page is not available then the request is redirected to the web server. The overall purpose is to reduce the retrieval time. Caching improves user perceptions about network performance in following two ways:

1) Caches hide wide-area network latencies by serving clients locally. The original content provider serves the client requests. The server load is balanced and its availability is improved by this way and hence the request serving time of server-side cache is reduced.
2) The network appears more reliable as the temporary unavailability of the network is hidden from users.

Proxy server creates a cache for each user session and then deletes at the end of the session. Whenever proxy listens to a client request, it checks whether the requested page is present in the cache, and, if it is present in cache then it reads the requested page from the cache and responds to client request by returning the requested page. If the page is not presented in the cache then it is requested from the main, server, and is sent to the client. Also, the page is added to the cache if cacheable. There are many techniques available for cache replacement policy. Clustering and classification are popular among them. The clustering is used to place similar websites into related groups and the classification is useful for predicting the cache lifetime and then adjusts the replacement priority to cache blocks. Web caching may have following drawbacks if the proxy is not properly updated as discussed in [19]:

- User may receive stale data.
- The origin server may become bottlenecks in case of growing number of users.
- The factors like system resources of cache servers (i.e., memory space, disk storage, I/O bandwidth, processing power, and net- working resources) may diminish the ideal effectiveness of Web caching.

The problem of updating such a huge collection of Web objects is still unmanageable even if the large caches are provided as solution. Hence, such an approach like web pre-fetching is needed, which will predict the future user's requests and retain in cache the most valuable objects.

B. Web Pre-fetching

Web Pre-fetching complements the web caching technique. It predicts the most likely web objects to be accessed by the user and fetches them into cache. The objective of doing so is to reduce latency. For those web objects, which are correctly predicted and fetched into the cache, the delay becomes almost zero [20].

C. Web Pre-fetching Techniques

Web pre-fetching mechanism predicts objects that are predicted to be accessed in near future, but users are not yet requested these. There are following techniques of web prefetching:

1) Interactive Prefetching Scheme
2) Link Prefetching
3) Top 10 Approach
4) Domain Top Approach
5) A keyword based semantic prefetching approach in internet news services
6) Dynamic web Prefetching
7) Web Comparison
8) Markov Model for Predicting web access

Web caching exploits the temporal locality and the Web prefetching schemes are based on the spatial locality of Web objects. The temporal locality may be referred as to repeated users' accesses to the same object within short time intervals, and the spatial refers to users' requests where accesses to some objects frequently entail accesses to certain other objects. The prefetching prevents bandwidth

underutilization and reduces the latency. It reduces bottlenecks and traffic jams on the Web are bypassed and objects are transferred faster. As a result, the proxies may effectively serve more users' requests, reducing the workload from the origin servers. Some problems posed during prefetching are as following:

- If the pre-fetched objects are not used then the prefetching increases the network bandwidth consumption.
- The server load is increased, by sending requests for all predicted objects that may not be subsequently used at all.
- The difficulty in predicting and prefetching the likely accessed objects increases as, most pages on any website are accessed only once and a lot of users access only one page from a particular website before moving on to a new website.

In order to overcome such problems, high accuracy prediction models have been used in [21].

D. Hybrid Approach

The Hybrid approach integrates techniques of web caching and web pre-fetching which may compliment each other as Web caching exploits the temporal locality and Web pre-fetching utilizes the spatial locality of web objects in logs. The temporal locality refers to repeated users' accesses to the same object within short time periods, whereas, the spatial locality refers to users' requests where accesses to some objects frequently entail accesses to certain other objects. In [19] a comparison has been performed on prefetching and caching and concluded that Web prefetching prevents bandwidth underutilization and reduces the latency. Table 1 represents the main difference between web caching and web prefetching

TABLE I
DIFFERENCE BETWEEN CACHING AND PREFETCHING

Approach	Locality	Architecture	Object's Placement
Web caching	Temporal	Pull based	Reactive
Web Prefetching	Spatial	Push based	Proactive

Web content can be effectively managed by exploitation of both the temporal and the spatial locality of objects.

E. Comparison of Web Mining Performance Improvement Techniques

Web Mining and Web caching are two effective solutions to lessen web service bottlenecks, reduce traffic over Internet and improve scalability of web system. But the hybrid approach is considered to be the best as it doubles the performance as compared to single caching. We have performed a comparative study of these techniques and the results are summarized in Table 2.

TABLE II
COMPARISON OF WEB MINING TECHNIQUES FOR PERFORMANCE IMPROVEMENT

Attributes	Web caching	Web Prefetching	Hybrid Approach
User Perceived Latency/ cache pollution Problem [16]	26% solved	Not solved	60% solved
Cache Hit ratio	Low	Moderate	High
Network bandwidth usage	Average	High	Low
Web servers' Load	Average	High	Low if carefully designed
Locality of objects	Temporal	Spatial	Temporal & spatial

IV. IMPROVED HYBRID APPROACH

The web prefetching requires anticipating future pages of users and preloading them into a cache. This means the web prefetching involves caching also. The Hybrid approach can be further improved by integration of Domain Top Prefetching and Dynamic Web Prefetching.

A. Domain Top Approach

Domain top approach for web prefetching combines the proxy's active the proxy's active knowledge of most popular domains and document [11]. In this approach proxy is responsible for calculating the most popular domains and most popular documents in those domains, then prepares a rank list for prefetching.

B. Dynamic Prefetch Approach

Dynamic web pre-fetching technique has been suggested in [12], by each user can keep a list of sites to access immediately called user's preference list, which is stored in proxy server's database. Intelligent agents are used for parsing the web page, monitoring the bandwidth usage and maintaining hash table, preference list and cache consistency. Prefetching is performed when web traffic is light and is reduced when it is high; hence, it controls the web 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. 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.

Our improved hybrid technique improves web caching and prefetching process by integrating both the Domain Top and Dynamic Pre-fetch techniques.

C. Benefits of Improved Hybrid Approach

Hybrid approach presents following advantages:

- It combines features of both web caching and web prefetching.
- It doubles the performance of web pages by prefetching the most frequently accessed objects into the web cache logs.
- Pre-fetching would now have wider scope as domain top approach can raise the hit ratio up to 20% in worst case and up to 450% in best case [11].
- The present of rank list (which maintains top domains and top documents) and preference list (which consists of list of sites stored by use to have immediate access) model will work well.
- This model will cover the loopholes like proxy server overheads.

V. CONCLUSION

The World Wide Web is today the major source of data and information for all domains. Web mining is an important and challenging activity that aims to discover new, relevant and reliable information and knowledge by investigating the web structure, its content and its usage. In this paper, we have presented a hybrid technique that combines the web caching and the web prefetching 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 pages. This paper has also presented the concept of optimized hybrid approach, which bring preference list from Dynamic technique into Domain top approach.

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. Cor fu 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] Alexandros Nanopoulos, Dimitrios Katsaros, and Yannis Manolopoulos, "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] Hendrik Blockeel, 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] Rudeekorn Soonthornsutee, Pramote Luenam," 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 Alexandros Nanopoulos, Dimitrios Katsaros, and Yannis Manolopoulos, "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] Josep Dome`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. Murali Bhaskaran," 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, Siti Mariyan Shamsuddin 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, Pokorny Jaroslav, "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.