Web Caching System by Using Hybrid Replacement Cache Algorithm (HRC)

Aung Ko Thet, Zaw Tun
University of Computer Studies, Yangon, Myanmar
aungkothetme@gmail.com, zawtun78@gmail.com

Abstract

Web is fundamentally a very simple idea. Web server that accept request from web Clients for pieces of information called web objects. When a user accesses a web page, the request typically goes out to the Internet to the server on which the page resides. The page content is then returned back across the Internet to the user's workstation. Web Caching works by storing the content of requested web pages on a dedicated caching server inside the network. The system would apply Hybrid Replacement Cache(HRC) Algorithm for caching another web pages when the cache is full. When the client makes a request, the browser is then checked whether it is in server cache or not and request information's frequency and last access time is recorded in the database. If it is in its sever cache after checking, the web information is displayed on the browser. Next step is considering how to cache the web objects and compare the size of the requested page and existing cache size whether the web object has to be saved in server cache or not. Applying this algorithm, this system reduce network traffic and user latency.

1. Introduction

In the last decade intensive growth of information volumes in the World-Wide-Web has led to the problem of network loads. Development of technical means for transmitting data and their introduction into practice do not correspond with the Internet growth. Hence, there is a need to find other approaches to solve the problem of network loads. The common method of solution without drawing additional technical means is via caching Web objects (text, image, etc.). Web documents are subdivided into two types: static and dynamic. Most of the Web objects are static documents which can be stored in certain data 'storage', i.e a cache, for the further usage. While reassessing these documents the proxy server checks whether there have been any modifications on the source site and if there have not been any, the user has the page downloaded right from the cache. Usage effectiveness of caching allows to use it on different levels such as Internet browser, proxy server of a local net and Internet proxy server. Proxies serving a large set of clients show the highest effectiveness of caching. It is connected with the fact that many Web users have correlation in requests (common interests).Researches have shown that 25-40% of all requested documents account for 70% of users' requests. Moreover, the nets of homogeneous organizations (university, corporation, etc.) have the highest correlation of requests. While caching Web objects on proxy servers as well as in other aspects of caching there appears a problem of finite storage of the cache. Because there is a need for a necessity of making room for new documents.

A replacement policy determines which object is to be removed from the cache. Selection of an effective policy can considerably increase caching effectiveness reducing network traffic by 20% and more. Difficulties in selecting a replacement policy are connected with specific characteristics of the network traffic which are the following:

-HTTP protocol gives access only to files of full size, i.e. a proxy server cache can satisfy the user's request only if the file has been stored completely (i.e. it can not store incomplete objects).So, the system cannot consider partial page caching.

-Documents stored in the proxy server cache are of different size, from several bytes to hundreds of megabytes.

-A stream of requests to the cache is a sum of streams of requests of hundreds and thousands of users.

Because sizes of documents are different there is a need to introduce a new metric of effectiveness called Byte Hit Rate apart from Hit Rate policy as the basic metric of effectiveness. Byte Hit Rate is computed as a ratio of amount of bytes derived from the proxy cache to the total amount of bytes requested by the user. Because of difference in documents' sizes metric suggested by Byte Hit Rate is the most adequate to determine the policy effectiveness as it is the one that shows the savings in network traffic.

The following section includes the Related work. Section 3 describes Background option: Web caching System. The proposed Scheme discussed in section 4. In section 5, we conclude the paper.

2. Related Work

B.D. Davison [3] pointed out three features

of Web caching make it attractive to all Web participants, including end users, network managers, and content creators. Caching reduces network bandwidth usage, which can save money for both content consumers and creators. It lessens user-perceived delays increases user perceived value and lightens loads on the origin servers. It is saving hardware and costs for content providers and providing consumers to get the shorter response time for non-cached resources.

C.W. Son and M. Busari [5] proposed for trace-driven simulations to evaluate the performance of different cache management techniques for multilevel Web proxy caching hierarchies. They considered heterogeneous cache replacement policies within a two level caching hierarchy, and size based partitioning across the levels of a caching hierarchy. Three different synthetic Web proxy workloads are used in the study, reflecting complete overlap, partial overlap, and no overlap in the workloads seen by the child-level proxies. The simulation results demonstrated that heterogeneous replacement policies and size-based partitioning each offer modest improvements in caching performance.

In paper [4], C.W. Son and G. Bai reported for the Web cache simulation step with two experimental factors: cache size, and cache replacement policy. The cache size determines the maximum number of Web content bytes that can be held in the cache at one time.

The cache replacement policy determines what objects to remove from the cache when more space is needed to store an incoming object. Six cache replacement policies are considered: removing objects at random (RAND), removing objects in the order in which they arrived (First-In-First-Out, FIFO), removing objects based on recent use (Least-Recently-Used, LRU), removing unpopular objects (Least-Frequently-Used, LFU), removing large objects (Greedy-Dual-Size, GDS) and Adaptive Replacement Cache (ARC).

A. Mahanti and C.W Son proposed that the common workload characteristics observed include a high degree of one-time referencing, a Zip-like document popularity distribution, heavy-tailed file and transfer size distributions, and a temporal locality property in the document referencing behavior [2, 6]. Among these characteristics, the slope of the Zip-like document popularity distribution is the most relevant to Web caching performance [2].

3. Background Option: Web Caching System

Caching has become a significant part of the Web's infrastructure. Caching has even spawned a new industry such as content delivery networks which are also growing at a fantastic rate. This system is familiar with relatively advanced Web caching topics such as the Internet Cache Protocol (ICP), invalidation, and interception proxies, so, now we would like to describe caching, explains how it applies to the Web, and describes when and why it is useful.

3.1 Caching Web Resources

Davison presented that web caching is similar to memory system caching. Web caching system stores web resources in anticipation of future requests. However, it will provides several web caching result from the non-uniformity of web object sizes, retrieval costs, and cache ability because of depending on the configuration of cache memory. To address object size, cache operators and designers track both the overall object hit rate (percentage of requests served from cache) and the overall byte hit rate (percentage of bytes served from cache) [3]. And also retrieval cost varies with object size, distance travelled, network congestion, and server load. Finally, some Web resources cannot or should not be cached, for example, if the resource is personalized to a particular client or is constantly updated. Caching is performed in various locations throughout the Web. These locations include two endpoints. They are Web browser and Web server.

Web caching system may be used because of popularity - the more popular a resource is, the more likely it is to be requested in the future. In one study [4] spanning more than a month, out of all the objects requested by individual users, on average close to 60 percent of those objects were requested more than once by the same user. Likewise, much content is the value to more than one user. In fact, the hits recorded in another caching study, up to 85 percent was the result of multiple users requesting the same objects.

C.W. Son and G. Bai explained that the caching effectiveness is traditionally measured by two quantities: the (document) hit ratio is the percentage of the total requests that are satisfied directly by documents stored in the cache and the byte hit ratio is the percentage of the total requested Web content bytes that are satisfied directly by documents stored in the cache. Both metrics are required since Web objects vary significantly in size. Other metrics such as user-perceived response time are dependent upon the hit ratio and the byte hit ratio, as well as network bandwidth, round-trip delay, and server load [4].

3.2. Potential Problem

This caching system has to face a number of problems associated with caching. First possibility is the end user seeing stale (that is, old or out-of-date) content, compared to fresh content

available on the origin server. HTTP does not ensure strong consistency and thus there is a real potential for data to be cached too long. Second, misses that are processed by a cache generally have decreased speed, as each system through which the transaction passes will increase the latency. Caching tends to improve the latency only for cached responses that are subsequently requested. So, web caching system can be able to increase the percentage of hit rate. Thus, a cache only benefits requests for content already stored in it. Caching is also limited by the frequency with which popular Web resources change, and, importantly, the fact that many resources will be requested only once. Finally, some responses cannot or should not be cached.

To overcome these potential problems, system must give expiration date far for all static content (buttons, graphics, audio and video files, and pages that rarely change) so that they can be cached for weeks or months at a time. By setting an expiration date far into the future, the content provider trades the potential of caching stale data for reduced bandwidth usage and improved user-perceived response time. A shorter expiration date reduces the chance that the user sees stale content, but increases the number of times that caches will need to validate the resource.

4. The Proposed Scheme

This section is going through the important stages of web caching system with clear explanation.

This system uses a cache for web storage. When the client makes a request, the request on the browser is then checked whether it is in its local cache or not, and the requested information is recorded in the database. If it is in its sever cache after checking, the web information is displayed on the browser. But, if the requested web object is not update in server cache, the sever cache is updated with the requested web object and display desire web page on the browser by directly using the internet. If the request is in Exception, it will not be sent to the server cache. If the request is not in Exception, it will be sent to the server cache and display on the browser by directly using the internet. The system would apply Hybrid Replacement cache (HRC) Algorithm for caching other adding web pages when the cache is full. Next step is considering how to cache the web objects and decide whether the web object has to be saved in the server cache or not. If the user is using any web object frequently, system cache this web object by using HRC.

Web Filtering Algorithm (For each request)

Let f= frequency of request

Let t= current time of request

```
begin
  while(cache is not full) do
    if ( url not include in Exception Table) then
       if ( url not include in URL Table )
            add this web site into the cache
                    f=1;
       else
            overwrite this web site into the cache
                     f=f+1;
       end if
    t=set the current time
  end if
end
```

The last step would describe the Least Recently Used algorithm. This algorithm removes web objects which have minimum requested frequency value (utility) and the earliest access time, so that popular web sites can remain in the cache.

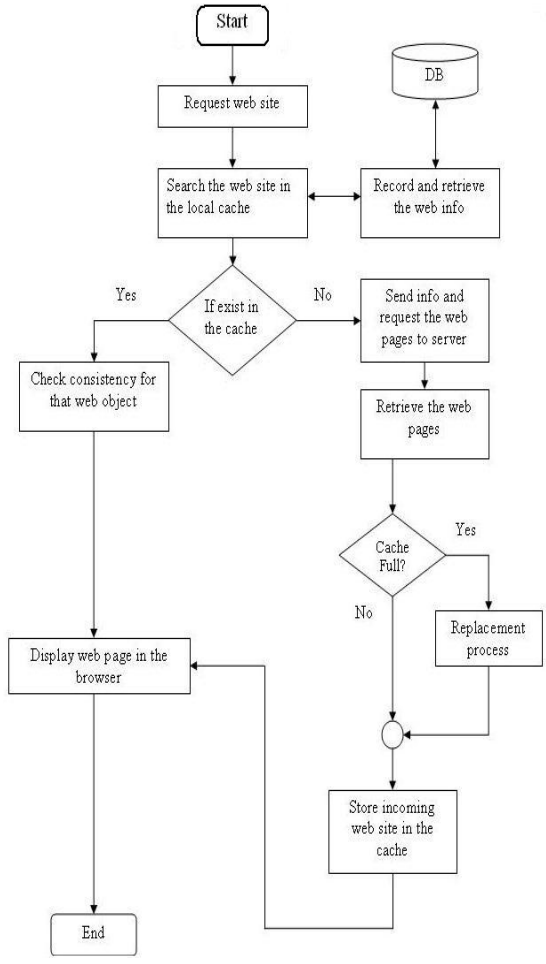

Figure.1: Proposed flow diagram

Adaptive Replacement Algorithm

Let f= frequency of request
Let et= earliest time

```
begin
    if ( cache is full ) then
       while(not end of cache) do
        begin
           search web site w which has min f
           if ( founded sites are more than one ) then
              begin
                    search et site
              end
          remove web site w from the cache
         end
end
```

Example of Operation

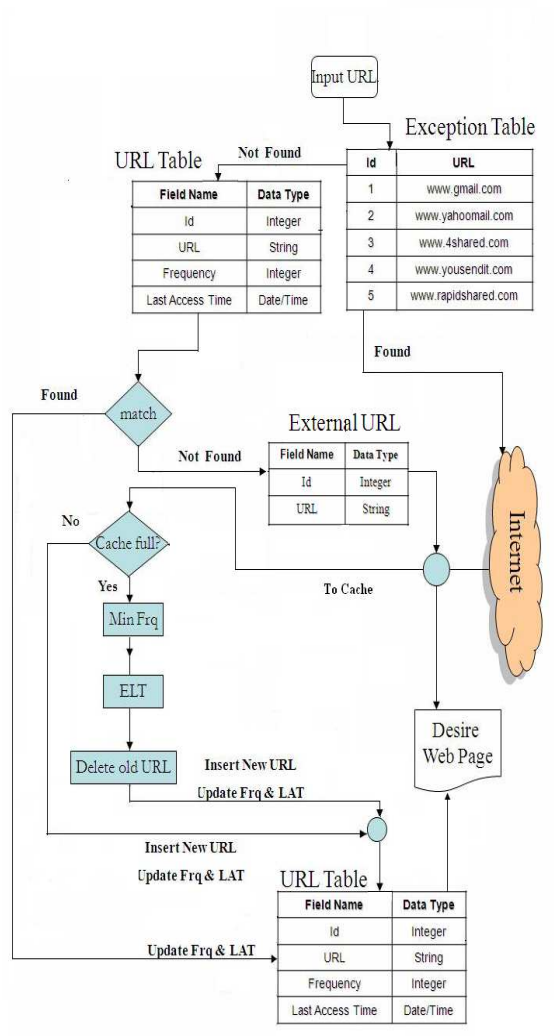

Figure.2: Example of Operation

4.1 Case Study

4.1.1 System performance by user latency

Base on testing, Web Caching System (WCS) reduces the traffic crossing the LAN by over 85%. It improves the data access performance significantly. For example, one result shows that over a connection with 100ms delay, a 2MB size web page that took 20 seconds to open normally but it took only 3 seconds with the web caching system. Web caching system reduces web page access times considerably.

In Figure.3, shows the user latency difference between using WCS and without WCS.

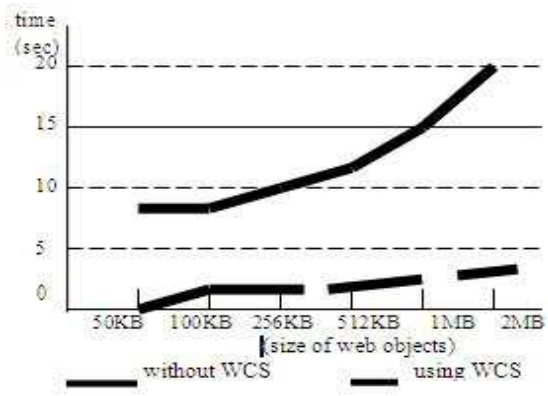

Figure.3: Reduction user latency

4.1.2. User environment

User environment is also important for measuring system performance. System performance also depends on the quantity of users. There will be two kinds of users (frequent user and guest user). User who frequently used this system is called frequent user. They will have same interest, inspirations or behaviors. They make convenient requests for the system. So, system can reduce storage space and system loads. If the quantity of frequent users increases, system can reduce computational cost and can give highest performance because of increasing hit rate.

User who sometimes uses this system is called guest user. It is impossible to concur with the web objects stored in the cache. If the quantity of guest users increases, calculation cost will be high and decrease hit rate. So, system performance will be lower than frequent user environment.

4.2 Benefit of the system

This HRC system can provide benefits to all parties on the web users, web proxy, ISPs, and original web servers. It can reduce user latency, bandwidth, network traffic and server loaded. When internet connection failed, this HRC system can continuously serve as online access by responding required web objects which had been stored in the cache. Therefore, users can access their requested web objects in the offline state. Since the whole page is cache in every request, user required data can access immediately and easy to implement the system.

5. Conclusion

As it is with any evolving technology, Web caching techniques are changing rapidly especially being of interest to both the Caches positioned near or maintained by the content provider improve access to a logical set of content. In this system we represented into effectiveness of HRC policy for the purposes of caching on a proxy server. The results of the experiments allow to make a conclusion about expediency of HRC algorithm on proxy servers in the Web. HRC algorithm cache only the currently active page of the sites but this page must not contain the private data entry (e.g For the page request with query parameters, or include some data items) to obey the privacy rule of the web information. For the page with query parameters, or include some data items (such as gmail inbox, bank account login field), will not be stored in the cache and this type of page must be added in the exception. As privacy of transactions becomes more important and harder and harder to achieve, it may be worthwhile to use caching proxies achieve anonymous access to what may be the most important achievement of the human civilization – written text information.

References

[1] A. Bestavros, A. Bradley, M. Crovella and P. Barford, Changes in Web client Access Patterns : Characteristics and Caching Implications, World Wide Web, Vol 2,No.1, January 1999,pp 15-18.

[2] A. Mahanti, C.W. Son and D.Eager, Traffic Analysis of a Web Proxy Caching Hierarchy, IEEE Network, Vol.14,No.3,pp. 16-23, May/June 2000.

[3] B.D. Davison, A Web Caching Primer, Department of Computer Science, Rutgers, The State University of New Jersey (USA), http://www.cs.rutgers.edu/davison/davison@cs.rutgers.edu

[4] C.W. Son, G. Bai, “Workload Characterization in Web Caching Hierarchies”, Department of Computer Science, University of Calgary {bai,carey}@cpsc.ucal.gary.ca.

[5] C.W. Son, M. Busari, Simulation Evaluation of a Heterogeneous Web Proxy Caching Hierarchy, Department of Computer Science, University of Saskatchewan.

[6] C.W. Son and M. Busari, On the Sensitivity of Web Proxy Cache Performance to Workload Characteristics, Proceedings of IEEE INFOCOM, pp.1225-1234, Anchorage, AL, April 2001.

[7] J.H. Bjornstad, Traffic Characteristics and Queuing Theory: Implications and Applications to Web Server Systems, Master Thesis, Department of Informatics, University of Oslo, May 22, 2006.

[8] P.D.J. Subhlok, Evaluation of Performance of Cooperative Web Caching with Web Polygraph, Department of Computer Science, University of Houston, Houston, TX77204, {pdu,jaspal}@uh.edu.