December 2001

Improved WWW Cache Updating For Rapidly Changing Objects

R. de Silva
University of New South Wales

Graham Low
University of New South Wales

W. Dewar
University of New South Wales

Improved WWW Cache Updating For Rapidly Changing Objects

R. G. de Silva, G. C. Low and W. J. Dewar
University of New South Wales
Sydney 2052, Australia

Abstract

In e-commerce applications and financial market services, information contained in the objects delivered by the web and database servers to the clients can rapidly change with time. Therefore, the challenge is not only to locate a cache that contains the required object but also to select the location of the most current version of the object. While there are few approaches to locate the cached objects in the Internet, extension of these methods to rapidly changing objects is not straightforward. As such, the researchers are attempting to provide a solution using active networking concepts. In this paper, we present a new method to improve cache updating in the Internet. In our approach, we calculate the mean of the update interval of the objects and group the objects into object groups with associated time slots. Then we use the time slot values to locate the best possible version of the object. We believe that our method would reduce the response time in obtaining the best updates of the objects and also reduce network traffic and cost.

Keywords: Internet caching, active networks, dynamic objects, e-commerce

1. Introduction

Web caching is a method used to store the contents of recently used web pages in an intermediate location and deliver the stored contents to the user if a subsequent query was made to the same URL. This has the advantage of reduced response time, network traffic and costs. However, with the growth of the Internet and e-commerce, the number of such caches in the Internet increases dramatically. As a result, much focus has been made during the recent years in developing efficient caching strategies. The Inter-cache Communication Protocol (ICP), for example, is a result of such research. While most objects in the web pages in the past did not change very often, the trend with the proliferation of e-commerce is to have web pages that contain objects such as stock in hand or price of a product to change more often. In some situations, such as in financial market systems, the frequency of changing an object or its value may be a few minutes or even seconds. With the introduction of active networking technology (Tennenhouse et al:1997; Tennenhouse & Wetherall: 1996; Wetherall et al: 1998; Legedza et al: 1998), where the network elements can participate in processing the information contained in the packets, web caching has to be seen in a different perspective. However, in the emerging active networking approaches, it is assumed that once an active node (AN) receives a request capsule from a client and if the capsule is redirected to the server, the response traverses the same path as the path from the client to the server in the opposite direction, so that the AN which received the original capsule, automatically receives the best update (Legedza et al: 1998). In the Internet, which uses TCP/IP, this AN may not lie in the backward path from the server to the client and as such, need not contain the most recent information in its cached objects. Thus, a more systematic approach is needed in updating and locating the cached objects and that is addressed in this paper.

In our approach**,** each object is assigned to a time slot and we use the time slot value of a group of objects as a measure of novelty, so that we can direct the request for locating the object to the cache that contains the statistically most recent version of the requested object. As the approach is statistical, our decision may not be the best, but it is better than the existing method of merely selecting one cache out of the caches that contain a copy of the object.

The rest of the paper is organised as follows. In Section 2 of the paper, we discuss the recent work on web caching and in Section 3, we introduce our approach for dynamic web caching. In Section 4, we conclude the paper and discuss possible future work.

2. Background

Web caching has the advantage of reduced response time, network traffic and the costs. As the popularity of e-businesses increases, e-commerce strategies proliferate the Internet and the need to cache and retrieve dynamic documents as well as the number of such caches will increase. In such a scenario, it is anticipated that dynamic price adjustment and distributed processing will dominate the e-commerce strategies. This situation already arises, for example, in e-commerce applications such as financial market services where stock prices can change rapidly. Rapid cache location and updating will be necessary to support this and will be a great challenge to achieve. As a result, much focus has been made during the recent years in developing efficient caching strategies. The Intercache Communication Protocol (ICP), for example, is a result of such research.

In an attempt to deal with this challenge, the principle of active networking has been used in the case of a server that supplies rapidly changing information to a set of clients (Legedza et al: 1998). In the active networking technology, one or more active nodes (ANs) within a data network can process the contents in the packets (Tennenhouse et al: 1997; Tennenhouse & Wetherall: 1996; Wetherall et al: 1998; Legedza et al: 1998). There may be any number of passive nodes in between two neighbouring ANs. The passive nodes do not respond to capsules but simply forward them to the next node in the path, thus, causing no problem. When a client needs to update an object, it sends a request capsule towards the server. When this capsule arrives at an intermediate AN, the capsule checks whether the requested object is available in the cache of the AN. If the object is available, the AN sends it to the client. If it is not available, the request is directed to the server. When the reply from the server traverses through the network, the object is cached in the ANs. Usually, it is assumed that the request capsule traverses the same path as the path from the server to the client (in the opposite direction), so that the AN which received the original capsule automatically receives the best update (Legedza et al:1998). In the Internet, which uses TCP/IP, this AN may not lie in the backward path from the server to the client and as such, need not contain the most recent information in its cached objects. Thus a more systematic approach is needed in updating and locating the cached objects that is addressed in this paper.

In legacy networks, there are five protocols for intercache communication but only ICP has the maturity and the longest history (Barish & Obraczke: 2000). In ICP, caches send queries to other caches to determine the best location from which it should retrieve the required

object. ICP consists of a request/reply paradigm (Wessels & Claffy: 1997) and is implemented with unicast over UDP. Hierarchical caching is described in the Harvest cache (Chankhuntod et al: 1996, Dutta: 1998; Gadde et al: 1997). In hierarchical caching, the caches are arranged in a tree-like hierarchy. When a parent cache does not have the required object, it can query the peer parent caches but not the child caches. If a child cache does not have the required object, it can query the peer child caches as well as its parent cache. In hierarchical caching, parent nodes can be overloaded with traffic during child query processing. It should also be noted that multicasting is the approach suggested in ICP to query peer caches.

Summary cache (Li et al 2000) attempts to reduce intercache communication during peer cache location (Fig 1). The proposal is to maintain a summary of URL contents of each peer cache in every other cache. Thus, when a caching node (e.g B) receives a URL request, if it cannot find the requested object in its own cache, it looks for the object in its summary caches. If the object is found in one of the summary caches (summary cache C in Fig 1), the request is directed to the node containing that cache. Since, not all URLs but a subset is stored, there is a possibility of directing the request to the originating server even when the requested object is available in one of the peer caches (Li et al 2000). A summary cache is not updated every time the cache of the peer is changed but periodically or when a certain percentage of the documents in the peer cache is not reflected in the summary.

3. Proposed Cache Updating

If the contents of the web pages are changing rapidly, the approach in summary cache will lead to a huge increase in the summary update resulting in heavy traffic flow between the peer caches. A second problem is that if the URL requested is available in more than one summary, there is no way of knowing the location of the most current object and a request has to be sent to all of them to select the latest version. Thus, for rapidly changing URL contents, we need to find a better mechanism for the retrieval of dynamic objects.

To alleviate the problem of increased traffic flow, we propose the following solution. Together with the requested object, the server sends its time stamp (STS). Using the present and the previous value of the STS, each AN constructs the update interval (UI) (i.e. the time difference between the present and previous values of STS) for each object that it caches. Each AN also stores the present and past values of the update interval (UI) of each object, and constructs the mean of it. The past values of UI can be stored in the hard drive and therefore, expensive DRAM space is not required. Depending on the mean UI value, each AN assigns the object into an object group that is associated with a time slot (TS). The duration of a TS can be few milliseconds, seconds, minutes, hours or days. When an AN multicasts the objects periodically to its peers, it sends the objects of the group together with the TS value. This reduces the traffic load as only one value per group is multicast. This multicast period is made much higher than the TS for that group. When an AN receives updated information about a cached object via a capsule, it updates its STS value and re-assigns the object to the correct TS group if the STS in the cache is less than the STS contained in the capsule. As time passes, the information contained in an object gradually becomes invalid. Therefore, each AN also uses its own UI statistics to determine the probability of validity of information (PVI) contained in an object. For example, if the

random variable UI is assumed to be belong to a normal distribution, PVI could be the probability that UI is greater than the current time minus the last update time point. Note that the current time and the last update time points are based on the clock of the AN that is attempting to update its cache and not on the clock of the originating server. When this PVI value falls below a certain threshold, the AN checks whether that object is stored in any other peer cache (by inspecting in its own memory), and requests the AN that has the lowest TS to send its cached information about the object. If this object's STS is greater than the cached object's STS, the AN replaces the object with the one received from the peer.

The proposed method can work in conjunction with any of the efficient WWW caching methods presently available such as summary cache (see Fig 2). When an AN receives a capsule from a client, it checks its cache and if the requested object is available, it will be delivered. If the, requested object is not available in its cache, it checks the summary caches and if it is available in more than one summary (say in C and D), the request is directed to the cache with lowest TS for that object. If it is not available in any of the caches, the request is directed to the originating server.

4. Conclusion

With the growth of the Internet and e-commerce, the number of web caches in the Internet increases dramatically. Therefore, developing efficient caching strategies is becoming more and more important. While most objects in the web pages in the past did not change very often, with the proliferation of e-commerce, important objects in a merchant's web page such as stock in hand, price of a product can change more often. In some situations, such as in financial market systems, the frequency of changing an object or its value may be a few minutes or even seconds. Although to address this issue an active networking approach has been proposed in the literature, our approach presented in this paper would complement and improve the cache updating.

Our method can be applied to active networks in conjunction with the exiting methods that are proposed for legacy networks. We believe that this method is very attractive if the contents of the objects are rapidly changing. The methods that are used in legacy networks do not have a similar feature so that if they were to be used for rapidly changing objects, every version of an object has to be treated as a different object. This will tremendously increase the intercache traffic as well as the memory requirements for the cached objects. We believe that our method would reduce the response time in obtaining the best updates of the objects and reduce the network traffic and costs. In our future work, we plan to investigate and compare the two cases via simulations. We also plan to investigate the influence of the mean UI, for example, by replacing it with a weighted moving average.

References

- Barish, G and Obraczke, K., "World Wide Web caching: trends and techniques", *IEEE Communications Magazine*, Vol 38, Issue 5, May 2000, pp 178–184.
- Chankhuntod, A., Danzig, P.B. and Neerdaels, C., "A Hierarchical Internet Object Cache", *Proc. USENIX*, 1996, pp 153-164.

- Dutta, P., "Internet object caching", *Proc. of the 7th IEEE Intelligent Network Workshop IN'98*, 1998, pp 95-118.
- Gadde, S., Rabinovich, M. and Chase, J., "Reduce, reuse, recycle: an approach to building large Internet caches", *The Sixth Workshop on Hot Topics in Operating Systems,* 1997, pp 93–98.
- Legedza, Ulana, Wetherall, David and Guttag, John, "Improving Performance of Distributed Applications Using Active Networks", *Proc. IEEE INFOCOM'98*, vol. 2, San Francisco, CA, March /April, 1998, pp 590-599.
- Li, Fan et al, "Summary cache: a scalable wide-area web cache sharing protocol", *IEEE/ACM Transactions on Networking*, Vol 8, Issue 3, June 2000, pp 281–293.
- Tennenhouse, D. L. and Wetherall, D. J., "Towards an Active Network Architecture", *ACM Computer Communication Review*, vol. 26, no. 2, April 1996, pp 5-18.
- Tennenhouse, D. L. et al, "A survey of Active Network Research", *IEEE Communications Magazine*, vol. 35, no 1, Jan 1997, pp 80-86.
- Wessels, D. and Claffy, K., "Internet Cache Protocol (ICPv2)", ftp://ftp.isi.edu/in-notes/rfc2186.txt, 1997.
- Wetherall, D., Legedza, U. and Guttag, J., "Introducing New Internet Services: Why and How", *IEEE Network*, vol 12, no 3, 1998, pp 12-19.

Figures

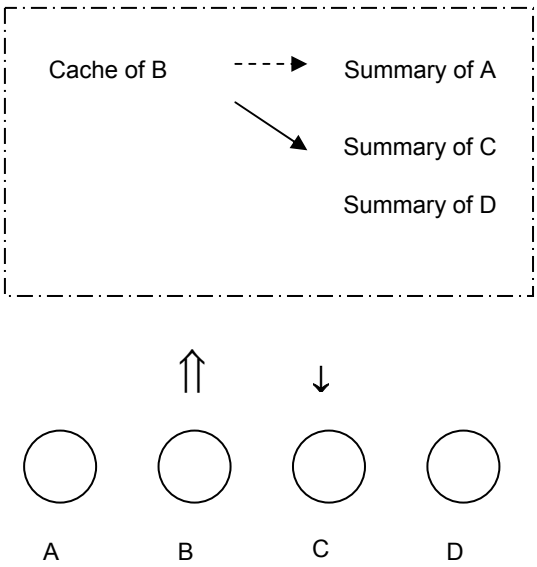

Fig 1: The object is selected from any cache which contains it

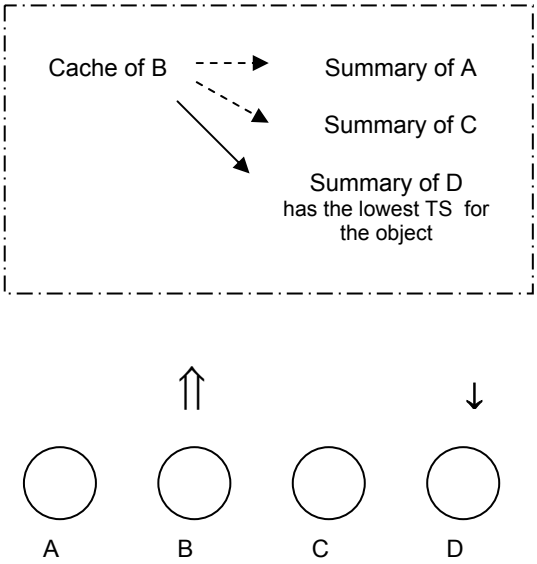

Fig 2: The object is selected from the cache with the lowest TS for that object