A Site-Based Proxy Cache

ZHU Jing (朱 晶), YANG GuangWen (杨广文), HU Min (胡 振) and SHEN MeiMing (沈美明)

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, P.R. China

E-mail: zhujing00@mails.tsinghua.edu.cn

Received March 14, 2002; revised September 10, 2002.

Abstract In traditional proxy caches, any visited page from any Web server is cached independently, ignoring connections between pages. And users still have to frequently visit indexing pages just for reaching useful informative ones, which causes significant waste of caching space and unnecessary Web traffic. In order to solve the above problem, this paper introduced a site graph model to describe WWW and a site-based replacement strategy has been built based on it. The concept of "access frequency" is developed for evaluating whether a Web page is worth being kept in caching space. On the basis of user's access history, auxiliary navigation information is provided to help him reach target pages more quickly. Performance test results have shown that the proposed proxy cache system can get higher hit ratio than traditional ones and can reduce user's access latency effectively.

Keywords proxy cache, site-based replacement strategy, access frequency, reconstruction technique

1 Introduction

Proxy cache has been proved to be an efficient means to alleviate Internet congestion, reduce user access latency and improve the scalability of WWW systems.

In traditional proxy caches, visited pages from any WWW server are cached in an indiscriminate way without considering the possible relation of pages in their original server. In fact, WWW is an enormous web with connections of any pages. Especially, pages in the same WWW server have some connections to each other. We are convinced that this organization nature of WWW is of great importance to an effective replacement strategy. In our research on proxy cache, a site graph model is introduced to describe WWW and a site-based replacement strategy is designed based on it.

In addition, we have noticed that when a user browses Web pages, he will often reach successive pages through links. Usually, user has to frequently access those indexing pages just for following the linkage from one page to another until he has found useful target pages. This will waste large amount of caching space, lead to unmeaningful hits and may well mislead the design of replacement strategy.

This phenomenon arises from the isolated page operation of the general caching system as well as the lack of interaction between the user and proxy cache. In this paper, a novel reconstruction technique is presented to solve this problem. With auxiliary linkage information appended in hit page copies, user can participate in the choice of page copies, and accelerate his access to target pages without the need to access extra indexing pages, thus actual access latency is largely reduced and effective usage of caching space is greatly promoted.

The rest of this paper is organized as follows. Section 2 gives a particular analysis of Web topology and Web users' access characteristics. Section 3 introduces the site graph model and site-based replacement strategy. Then our novel reconstruction technique is explained in detail in Section 4. The final section presents the performance test results of our proxy cache system.

2 Web Topology and Web Access Characteristics

2.1 Net-Like Topology Structure of Web

Web can be considered as a multi-server/multi-client net-like topology. This can be attributed to the interlinkage of Internet as well as the HTML language used to describe Web pages. As we know, the Anchor element is the essence of HTML. The text between the opening and closing anchors constitutes a hypertext link (such as (A

This work is supported by the National Research Foundation for the Doctoral Program of the Higher Education of China (No.99000309) and the National Natural Science Foundation of China under Grant No.60173007.
HREF="http://www.tsinghua.edu.cn/"/A}. If user selects the link, he will move from the current page to the document specified by the value of the link. So there exist connections between Web pages. In particular, inside a single Web site, the contents of different Web pages are more relevant to each other, and it is also necessary for these pages to reference one another so as to facilitate Web users' access behavior.

### 2.2 Web Access Characteristics

Based on the analysis of Web access traces and other log files, studies have revealed the following characteristics of Web access: heavy-tailed distribution of Web pages in Web servers; Zipf distribution of HTTP requests; self-similarity of Web traffic[3-6]. These characteristics indicate temporal locality and reference locality of Web traffic, which laid the foundation of the feasibility and effectiveness of proxy cache technique. But these studies of Web traffic and HTTP requests have neglected such a fact that users' access to Web pages is mainly accomplished through the linkage between Web pages. This simple attribute has significant implication for designing an effective proxy cache.

In experience, we can surely find that in any given period, the user's requests are usually on the same Web server. Due to the complexity of URL used to identify Web pages, it is nearly impossible for a user to visit each page by inputting its URL in the address window of Web browser. In most cases, Web user begins his access from the entry page of a Web site by putting in the URL of that site, and then follows the linkage between pages until he reaches the needed pages. So, the user's access behavior can be described as a sequence of requests. For example, if the user begins his access from page $p_i$, and then to $p_{i_0}, p_{i_1}, ..., p_{i_n}$, until $p_j$, then $p_i, p_{i_0}, p_{i_1}, ..., p_{i_n}, p_j$ constitute his access sequence. Generally, such pages as $p_{i_0}, p_{i_1}, ..., p_{i_n}$ only provide entrance guidelines to help user find his destination, so they are usually called indexing pages; while those which provide real useful information for the user (such as $p_j$) are called target (informative) pages. In fact, any Web pages can be classified into these two categories.

When analyzing such access process, we find that if user's access destination is the next Web page instead of the current one, but the linkage to the desired page is only given in the HTML code of the current page, then the user has to access the current page too. And it has been demonstrated that access to such indexing pages accounts for a large proportion of total Web requests[7].

### 2.3 Implications for Proxy Cache Designing

What hint can we expect from the net-like Web topology and the characteristic of user's access behavior? Two main issues are discussed here.

1. When the requested Web pages are cached, the linkage between pages in original servers can also be saved. This will be beneficial for the storage and search of page copies in proxy cache server. More importantly, the connections between pages will serve as a significant factor in determining a Web page's relative importance in replacement strategy. This is what our site-based replacement strategy features in.

2. Imagine such an ideal scene: suppose the user has followed links $p_i, p_{i_0}, p_{i_1}, ..., p_{i_n}$ to reach $p_j$ (target page). When he hits in next visit to $p_i$, the linkage to $p_j$ immediately appears in the page copy $p_i$ returned to him, offering him the option to access $p_j$ directly without the need to visit $p_{i_0}, p_{i_1}, ..., p_{i_n}$. So if an auxiliary linkage shortcut exists from the entry page to the destined page, user's access latency will be greatly reduced. This is the original intention of our reconstruction technique.

### 3 Site-Graph Model and Site-Based Replacement Strategy

According to the connections between Web pages, a Web site in WWW can be described as a site graph $W=(P,E)$, whose vertexes ($p_i \in P$, $i=0,\ldots,s$) are pages saved in the Web site, and $p_0$ is the entry page of that site. Directional edge ($e_{ij} \in E$, $i,j=0\ldots s$) means that one can access page $p_j$ from $p_i$. For a vertex $p_j$, all vertexes that have edges to it constitute its precedent vertex set $PVS(p_j)$. And all vertexes that have edges from $p_j$ form its succeeding vertex set $SVS(p_j)$. The direct succeeding vertex set is denoted as $SVS_1$, the direct succeeding vertex set of $SVS_1$ is $SVS_2$, and so on. It is the same for PVS.

With the help of the site graph model defined above, we can trace and analyze user's request conduction. When the proxy server receives a user's request, it generates the path matrix of the corresponding Web site if it is requested for the first time. At that moment, the Web site graph in-