Proxy 和 Cache 分析

王春雷　龚俭*

(东南大学计算机系　南京 210096)

摘要 ：本文介绍了有关 proxy 和 cache 的基本概念和工作方式，以及目前流行的软件 Netscape Proxy Server 的使用。通过分析华东(北)地区网络中心 cache 的使用情况，对在 CERNET 推广使用这项技术提出建议。

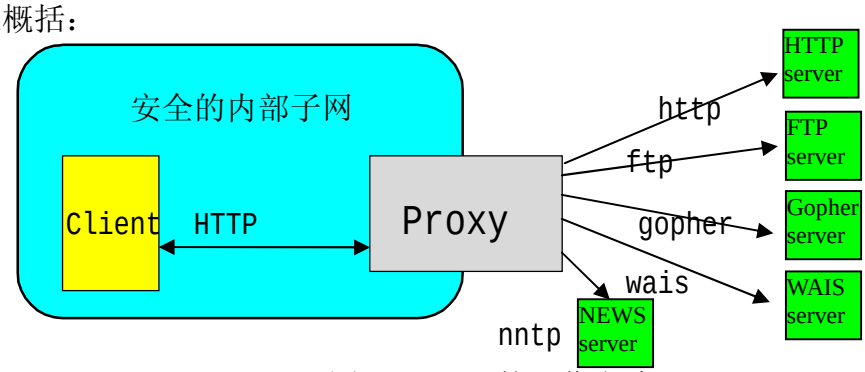
关键字 ：proxy，cache，访问控制，过滤

一 前言

随着 CERNET 的发展，越来越多的人能够连上 Internet 网，对世界各地的网点进行访问。由于国内的信息资源比较有限，许多信息需要从美国或其他国家获取，在目前的条件下，远距离传输的延时和可靠性是不令人满意的，而且同一个信息重复传输，造成时间和资源的浪费。另外，流量控制和安全因素也是人们关心的问题，因此“网络代理”(proxy)应运而生。

二 proxy 和 cache 工作原理

最初的 proxy 是为了让客户机能够从防火墙内访问外部资源而设计的，防火墙挡住了内部子网和外部的通信，而 proxy 则在另外一个地方提供服务，当然为了保证安全，它需要比较严格的验证、加密等手段。Proxy 的功能可以由下面的图来概括：



图一 proxy 的工作方式

各种协议都可以有自己的 proxy，其中用得最多的是 WWW 的代理，下面的介绍和统计都是针对 WWW 的。作为子网内的所有客户对外访问的中继，Proxy 可以获取流过的所有信息，因此人们为它加入了“缓冲”(cache)功能，把经过 proxy 的资源存储在本地的 cache 中，当子网内的用户再次请求时不必从原主机传输，而是使用 cache 中的内容，这样既节约了远程传输的网络开销，又减少用户的等待时间。

早期的 proxy 软件是由用户自己开发的，它首先具有一个 Web Server 的功能，另外还增加了一些额外的处理、存储、转发、加密等功能。经过不断改进，proxy 的算法和工具已经发展到了很高水平，一般 Proxy 包含这些功能：

1.访问控制：限制某些用户或某些主机可以使用 proxy 的服务；

2.过滤：出于安全的考虑，限制用户访问的 Web 服务器地址或内容；

* 王春雷，东南大学计算机系研究生，研究方向：计算机网络、分布式处理；

龚俭，东南大学教授，CERNET 专家委员会成员，研究方向：开放分布式处理、网络管理、网络安全。

3.加密：Proxy 可以为用户和主机之间的通信进行加密，保证信息传输的安全性；
4.cache：节省用户时间和通信资源，减轻主机的负荷。

Proxy 给用户带来了明显的好处，所以已经得到广泛的使用。比较著名的软件有 Netscape Proxy Server[3]、Microsoft Proxy Server，著名的系统包括 Harvest[4]、UK National Web Cache at HENSA Unix [5]等。

三 Netscape Proxy Server 的使用要点

Netscape Proxy Server 是由 Netscape 公司开发的一个商业软件，是目前较常用的一种，它向教育部门提供免费版本，与其他 proxy 服务软件相比，它具有以下特点：

1，用户界面良好，Netscape proxy server 的安装、维护、配置等功能都是集成在浏览器界面上，用户可以在任何平台的主机上对 proxy 进行远程管理。
2，良好的安全功能，它支持 SSL(安全的 Socket 层)标准。
3，优良的性能，Netscape Proxy Server 支持大容量的 Cache 库(64GB)，并采用 RSA MD5 方法进行资源定位，查询速度非常快。
4，功能全面，支持 PASV(反向的 proxy)，SOCKS 等增强功能。并在管理界面中提供完整的工具包，使用户可以很方便地统计、分析日志，监视当前的运行。

在安装完成后，proxy 将在两个口上提供服务，一个是提供管理服务，它是一个标准的 WWW 服务器，另一个是 proxy 服务。Proxy 通过管理界面提供系统设置、路由、URL 表、访问控制、加密、日志等九类功能。

访问控制的目的是限制使用 proxy 的用户，有基于用户验证和基于主机两种方法，基于用户是在 proxy 上维护一个用户库，每次访问都会进行口令验证，然后决定是否允许访问；基于主机的控制方法是：用户必须在指定的主机上，才能得到 proxy 的服务。

过滤是对用户访问的服务器的限制，它也支持多种过滤方法，比如对文件格式(MIME)进行控制，拒绝某些格式的文件被访问；可以过滤 HTML 的标记，象<applet>、<object>等；另外还可以对用户使用的浏览器或用户请求进行过滤。当然最常用的过滤方法是对地址和域名的过滤，在 Filter | URL Filters 界面中，可以定义两张表，一张为允许访问的 URL 表，另一张为不允许访问的 URL 列表。一个 URL 必须不在 Deny 表中且在 Allow 表中才能被访问，所以一般两者中之一置为 NONE。Netscape proxy 定义了一种正规表达式语法，使管理员可以在 URL 中使用通配符，具体定义如下：

1） 。 ：句号匹配任意的单个字符；　　2） * ：长度任意的字符串；
3） 。*：长度大于零的任意字符串；　　4） \。 ：表示句号本身
5）[a-t]：在 a-t 之间的任意单个字符；　6） (a | b | c) ： a,b,c 三者之一
7）[^ 0 1 2]：除 0,1,2 之外的任意单个字符；
8）*(.....) ：除括号内内容外的所有长度大于零的字符串

例如，假如只允许用户访问国内的 WWW 服务器，只需要把 allow 表中定义图二所示。

```
http://.*\.cn/.*                    #表示所有形如 http://*.cn/* 的 URL
```

图二　允许访问国内地址的登记表

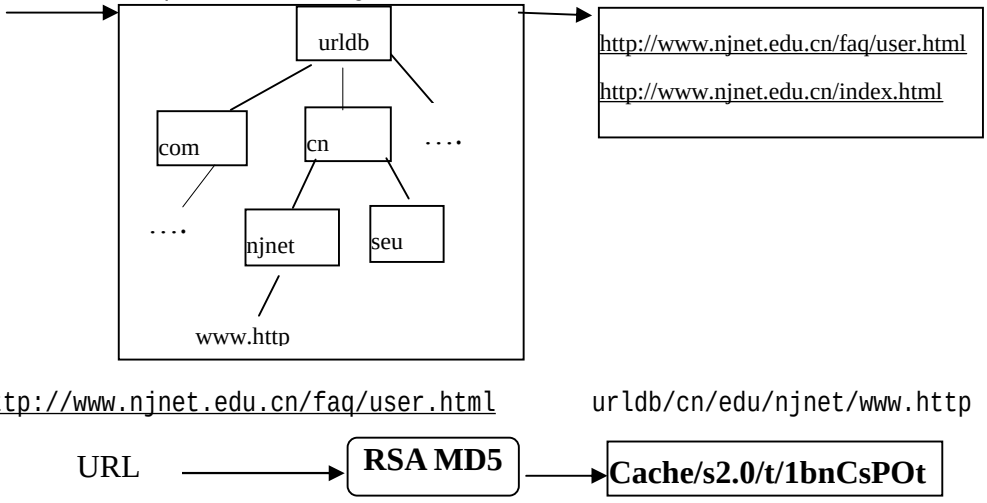
应该注意的是，图二定义的表允许用户通过 DNS 系统访问服务器，而直接使用 IP 未必能行。在安装 proxy 时有一个选项“是否每次访问都做 IP 和 DNS 的转

换？”，管理员如果把它设置为“否”，则在上面的例子中，当用户使用 IP 地址访问时将全部被拒绝，如果设为“是”，那么系统先进行 IP 和 DNS 的转换，然后再判断是否允许访问，而那些没有 DNS 的服务器还是会被拒绝。

四　cache 使用分析

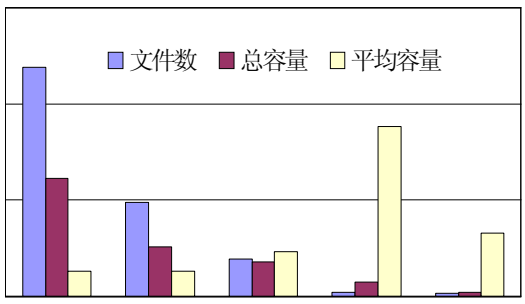
华东(北)地区网络中心采用了 Netscape 公司为教育部门用户开发的 Proxy Server，通过几个月的使用，已经积累了大量的 cache 和日志。目前它只为网络中心内部用户提供 WWW 的 proxy 服务，用户数为 10 – 20 个， cache 容量定为 500M，目前它保存了 7000 多个文件，cache 总容量达到 60M。Netscape proxy server 可以按用户的要求记录详细的日志，还提供工具对日志进行分析处理、备份等工作。

cache 系统把每个资源保存为一个文件，这些文件分布到不同的目录上，Netscape Proxy Server 提供分布算法保证每个目录上的文件数量大致相近。资源的文件名是由它的 URL 通过 RSA MD5 算法计算出的一个 8 位字符串。安装 proxy 时管理员可以设置把 cache 中能查到的 URL 都记录在一个目录中(与 Internet 域名系统相似的层次型的数据库，一般名为 urldb，如图三)。管理员可以通过界面中的 Cache | Cache Management 管理目前积累的所有 Cache 文件。

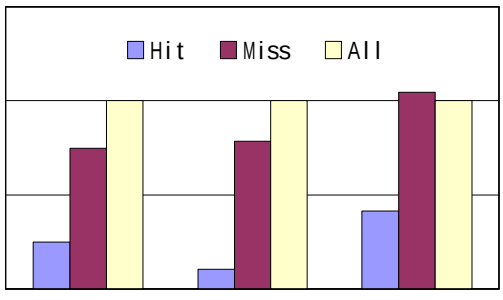


图三 URL 与 Cache 文件名的转换关系

通过对 cache 文件和 proxy 访问日志进行分析，我们发现下面的一些结果：



图四 cache 文件格式分布



图五 cache 文件的命中率

首先，Web 传输的文件格式非常集中：图四从左到右分别为五种文件格式的统计：image/gif、text/html、image/jpeg、application/octec-stream、text/plain ,三个直方块表示文件数，文件总长度和文件平均长度。图中显示，

在 MIME 定义的上百种不同文件格式中， Web 最常用的有 gif、html、jpeg 三种，它们占文件数量和流量的绝大部分，基本是长度相近的小文件；二进制文件单个长度较大，主要原因是许多软件的 download 也通过 http 协议进行。其它文件只占很小部分，这些结果和目前 WWW 的现状相符。

第二，访问的网点非常集中，从统计结果看，其访问范围是集中于一百多台主机(超过 20 个文件被 cache)，这说明用户的访问内容很集中，图七的曲线符合文献[1]中提出的 Zipf 规律。

第三，cache 的命中率不高：访问日志中记录的响应有六种

(1) NO-CHECK，在 cache 中找到文件；

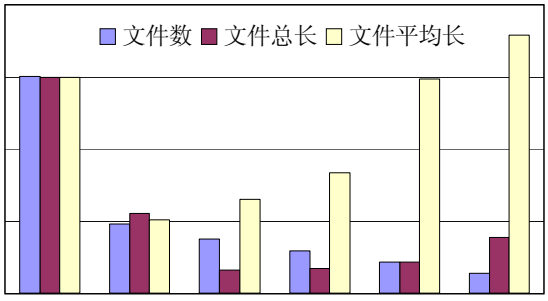
(2) WRITTEN，写入 cache;

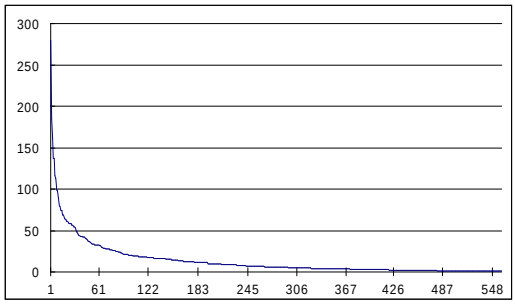
(3) REFRESHED，刷新，当 cache 中文件过于陈旧时，直接对原服务器进行访问；

(4) DO-NOT-CACHE，文件的作者希望用户能经常访问原文，或 proxy 设置为对某些网点(如本地网点)不做 cache；

(5) NON-CACHEABLE，无法 cache，这一般是对网络上的执行文件，如 CGI 等，它的返回跟输入参数有关，保存执行结果没有意义；

(6) 其他情况，如中断、文件未传完等；





图六 proxy 对用户请求的响应分类　　图七 节点的访问频率统计

图六的统计表明，从得到响应的数量上看，no-check(在本地 cache 中得到响应)排在第二位，占 25%，比著名的 cache server 低 (一般 cache 对请求的命中率为 30% -50%)。但是它的文件平均长度比较小(是普通文件的 40%)，所以在本地响应的流量只占总流量 11%。我们认为使用者较少影响了命中率，且短时间内 proxy 工作还不稳定。

总结上面的统计数据，我们发现 Web 访问中大量无法 cache 的信息，如 cgi 请求，相对削弱了 cache 的作用，图六显示，它占请求数的 20%和流量的 11%，cache 只能对静态的信息起作用，所以影响了它的命中率。另外在使用过程中我们还发现：

(1)信息定期更新没有很大意义，设置有效期(expire)的目的是保持资源的内容“新鲜” ,但是根据统计，文件需要“刷新”(refresh)的概率非常小 (0.4%)，

资源的所有者很少修改文件，我们不妨把文件的有效期定得长一点。

(2)cache 的容量和置换算法会影响系统的效率，不过为每个用户保留 20—50M 空间一般足够使用。根据有关统计资料[1]表明，除了一些用户很多的公用 cache server 外，很少发生 cache 满的情况。

(3)用户的响应时间：有了 cache 后，不管是否命中，用户请求都会在 cache 中查找，当 cache 容量增大时，查找时间还会延长， Netscape proxy server 的响应延时主要由于它采用的 hash 方法有可能会“碰撞”。当然这些时间跟用户需要远程传输相比是非常小的，事实上在使用中，用户基本感觉不到 proxy 和 cache 的存在。

五，结论

Proxy 软件功能可以分为两个部分：一部分负责访问控制、过滤、加密等功能，另一部分负责 cache 的维护。前者用于对用户的管理，后者用于用户服务。两者都工作在用户和服务器之间，通常结合起来使用。根据在华东(北)网络中心使用 proxy 的经验和上面的统计结果，结合目前 CERNET 的现状，我们建议在 CERNET 内部大力推广 proxy 和 cache 技术，这是投资小见效快的好方法。

使用 proxy 必须有一定的条件，服务的范围过大或过小都会降低效率。**如果主要是使用它的管理功能，它适用于一个用户较少、有防火墙的比较封闭的内部网环境**；如果为所有用户提供公开的服务，则主要使用其 cache 功能。Proxy 和 cache 都应靠近用户，否则就不能达到节省时间和资源的目的；使用的用户越多，cache 容量越大，命中率会越高，但同时对 cache 服务器的要求也提高。

目前世界上比较著名的 cache server，如 Harvest[4]，UK National Web Cache[5] 等，可以为更广范围内的用户服务，而且有很高的命中率(约 60%)。它们一般采用高性能的服务器或服务器阵列、快速通信器材、大容量磁盘等设备，能够同时处理大量用户请求。而且使用自己开发的服务软件，对 cache 置换、用户响应、文件匹配进行优化，能够显著地提高网络的使用效率。在积累大量 cache 文件的基础上还可以开发新的应用，如查询、目录、排序等。在 CERNET 内部，由于范围较大，建议在地区网络中心设立公用的 cache server，各用户可以就近使用，并且在多个 cache 之间加强合作(Internet 已有 ICP -- Internet Cache Protocol-- 支持这种应用[6])，提高网络的使用效率，节省用户的费用和时间。各个学校或部门，可以根据自己的需要建立具有 proxy 和 cache 功能的服务器，对用户的访问进行管理。

[参考文献]

1. A caching relay for the World Wide Web , Steven Glassman , Computer Network & ISDN Systems 27 (1994) 165-173

2. World-Wide Web proxies , Ari Luotonen , kevin Altis , Computer Network & ISDN Systems 27 (1994) 147-154
3. http://www.netscape.com/comprod/server_central/product/proxy/index.html
4. http://harvest.cs.colorado.edu
5. http://www.hensa.ac.uk/wwwcache/
6. http://excalibur.usc.edu/icpdoc/icp.html

Proxy and Cache Analysis

WANG Chunlei GONG Jian

Department of Computer Science and Engineering

Southeast University , Nanjing 210096

Abstract: The paper introduces the fundamental principles and mechanism of Proxy and Cache , as well as the Netscape Proxy Server , a leading software in this field. After the data of Eastern China (North) Center cache server analyzed the proposal of spreading this technology in CERNET is also launched.

Key Words: proxy, cache, access control , filter