基于 Web 日志分析的 Web QoS 研究

田昌鹏
（重庆工商大学　重庆 400067）

摘　要　Internet 的服务模式正由传统的数据通信与信息浏览向电子交易与服务转变，需要对不同的用户或 HTTP 请求提供区分服务和性能保证。本文通过对网络 Web 服务器日志中记录的大量用户信息进行 Web 日志分析，研究在 Web 服务器中及其系统中引入和实现 Qos 控制的机制和策略，了解用户习惯，提供个性服务，提高服务质量和效率。

关键词　Web 挖掘，日志分析，Web QoS

Based on the Analysising and Researching Webserver Log of Web QoS

TIAN Chang -Peng
(Chongqing Technology and Business University, Chongqing, 400067)

Abstract　The Webserver log has the records of much information of users, making analysis on the log is beneficial to webmaster to obtain users' habits, the information needed by the users, and improve service quanlity and efficiency, provide individual service. This paper introduces Web Minning technology, Web QoS technology ,Webserver log analysis, format of Webserver log, theory of log analysis, analysis tools used and summarizes our practical experience of analysis on Webserver log.

Keywords　Web mining,Log analysis, Web QoS

1　引言

根据 2007 年 1 月 23 日，中国互联网络信息中心(CNNIC)发布的《第 19 次中国互联网络发展状况统计报告》显示，截至 2006 年底，我国网民人数达到了 1.37 亿，目前 Web 流量成为 Internet 上信息传输的主流，全国网页数和网页字节总数分别为 44.7 亿个和 122,306GB，与去年同期相比分别增长 86.3%和 81.7%[1]。随着信息技术的迅速发展，网络上的 Web 应用以及 HTTP 请求也迅猛增长，Internet 的服务模式正由传统的数据通信与信息浏览向电子交易与服务转变，由此而来的问题是如何为用户提供满意的服务性能保证。

随着 Web 应用和电子商务应用的发展，大学、企业和服务商都越来越推崇将传统的产品、管理和服务转移到 Web 上去，如：E-learning、网上购物、在线银行、在线股票交易等，门户网站每天都有大量用户对它进行浏览、查询、下载等访问，访问信息都被记录在了 Web 服务器的日志文件中。通过对 Web 服务器日志信息进行统计分析，能够有效地了解用户的行为和偏好，从而加强对网站及其内容的维护和管理，提高服务的质量和效率。

近年来，网络传输中的 QoS 技术研究十分活跃，伴随着网络 QoS 技术研究和应用的不断发展，一种面向 Web 客户和 HTTP 请求提供性能保证及服务区分的技术—Web QoS 应运而生，并且在国际上得到越来越多的学者和业界商家的瞩目，成为 QoS 技术的一个新的研究领域和重要的学术分支。但目前通用的 Web 服务器尚未支持 Web QoS 机制，无法为 Web 应用提供服务区分和性能保证。因此，如何在 Web 服务器中及其系统中引入和实现 QoS 控制的机制和策略，成了实现下一代网络 QoS 控制技术不可或缺的环节。

2　Web 日志分析及其应用

2.1　Web 日志

对一个电子商务网站，其完整的 Web 服务模式如图 1 所示。浏览器发送一个 Portal 页面的请求；Portal 服务器解析用户的请求，并将认证和权限请求递交给身份认证服务器 IDS；IDS 认证用户，返回结果给门户服务器；门户页面内的各个 Portlet 节点收到请求后分别产生页面片断(这里可能调用 App 服务器及数据库)，如果该页面 Portlet 需要验证，Portal 节点就会对登录服务器进行验证；Portal 节点包含页面定义、显示页面的代码和页面用户的定制；门户服务器负责将用户的内容进行组装，统一响应到用户浏览器[2]；Web 日志服务器负责记录下 Web 服务器接收处理请求以及运行时错误等各种原始信息。

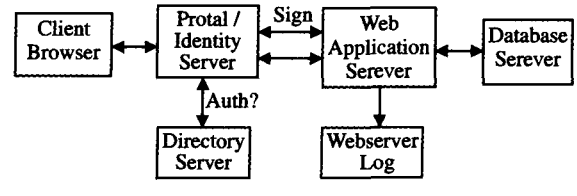

图 1　电子商务网站信息流示意图

随着 WWW 上可用信息资源的爆炸性增长，把数据挖掘和 WWW 这两个领域结合起来进行“Web 挖掘技术”研究，一方面能帮助用户定位、分析和评价所需的信息，另一方面能帮助服务提供者追踪和分析用户的访问模式，以利于提供更好

田昌鹏　硕士生，主要从事计算机网络应用研究。

的服务。Web挖掘可分为Web内容挖掘、Web结构挖掘和Web日志分析三类。

Web日志分析就是从Web服务器所产生的庞大的日志文件中，挖掘出隐含的、有用的、尚未发现的信息和知识，经过分析加工所得到能直观地被用户看懂的、有价值的信息和知识的各种分析结果。Web日志分析被认为是目前解决网站"数据丰富、信息贫乏"的一种有效方法。通过各种分析报告，网站的运营者可以了解用户的行为和偏好，了解信息内容的受关注程度，了解网站可能存在的问题，促进其对网站及其内容的维护和管理，提高服务的质量和效率，提供资源建设的决策支持等[3]。

2.2 Web日志分析的实现原理

网站服务器日志记录了Web服务器接收处理请求以及运行时错误等各种原始信息，如图2。通过对日志进行统计、分析、综合，就能有效地掌握服务器的运行状况，发现和排除错误原因、了解客户访问分布等，更好地加强系统的维护和管理。Web服务模式主要有以下3个步骤：(1)服务请求。客户端通过浏览器向Web服务器发出请求(如get)，根据HTTP协议，这个请求包含了客户端的IP地址、浏览器的类型、请求的URL等一系列信息。(2)服务响应。Web服务器接收到请求后，根据请求将客户端要求的信息内容返回到客户端。如果出现错误，那么返回错误代码。(3)保存信息。服务器端将用户访问信息记录到日志文件中[4]。(如图1)

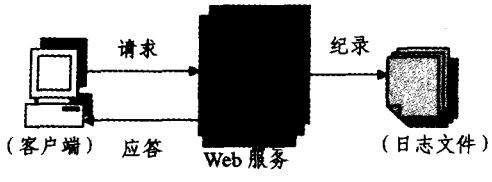

图2 Web服务模式示意图

2.3 Web日志分析

下面是标准的一段日志文件格式示例：

211.83.196.238[10/Dec/2006:13:40:44 +0800] "GET /2007/js.js HTTP/1.1" 200 2588 "http://www.ctbu.edu.cn/2007/main.php" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"

10.15.33.25[10/Dec/2006:13:40:44 +0800] "GET /2007/images/006.jpg HTTP/1.1" 304 - "http://www.ctbu.edu.cn/2007/" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; TencentTraveler)"

10.15.32.17[10/Dec/2006:13:40:44 +0800] "GET /2007/ HTTP/1.0" 200 11599 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"

219.221.34.101[10/Dec/2006:13:40:43 +0800] "GET /2007/images/login.jpg HTTP/1.1" 200 20063 "http://www.ctbu.edu.cn/" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"

通过日志示例及Web服务器日志格式(见表1)我们可以看到Web访问日志记录了服务器接受请求以及运行状态的各种原始信息，包括客户端的IP地址、访问发生的时间、访问请求的页面、Web服务器对于该请求返回的状态信息、返回给客户端的内容的大小(以字节为单位)、该请求的引用地址、客户浏览器类型等。Web日志还包括了上次访问页面、cookies等更多我们迫切需要的信息。如果客户端在连续的网站上浏览，就会产生很多条包含某些相同信息的日志文件。这样就构成了一个访客在网站上所有活动的日志信息列。通过这些信息，我们就可以通过一定的算法，从而了解到这个访客在网站上的整体行为。通过对这些信息的统计、分析和综合，就可以识别用户，了解访问分布，掌握服务器的运行状况等。

表1 Web服务器日志格式

域(field)	描述(description)
日期 (date)	请求页面的时间、日期和时区(date, time and timezone of request)
客户端IP (client IP)	远程主机的IP或DNS入口(remote host IP and/or DNS entry)
用户名 (username)	远程登录的用户名(remote logname of the user)
字节 (bytes)	发送和接收的字节(bytes transferred sent and received)
服务器 (server)	服务器名称、IP地址和端口(server name, IP address and port)
请求 (request)	URL查询和枝节(URL query and stem)
状态 (status)	返回给HTTP状态标识(http status code returned to the client)
服务名 (service name)	用户请求的服务名称(requested service name)
耗用时间 (time taken)	完成浏览的时间(time taken for transaction to complete)
协议版本 (protocol version)	传输用的协议版本(version of used transfer protocol)
用户代理 (user agent)	服务提供者(service provider)
Cookie	标识号(cookie ID)
参照页 (referrer)	本页的上一页(previous page)

3 Web QoS及问题

3.1 Web QoS的概念

QoS是指网络在传输数据流时要求满足的一系列服务请求，强调端到端或网络边界到边界的整体性，具体可以量化为带宽、延迟、延迟抖动、丢失率、吞吐量等性能指标。QoS控制技术的基本目标是为Internet应用提供性能保证和服务区分。为此，IETF(Internet Engineering Task Force)已经提出了两种不同的Internet QoS体系结构，即综合服务(Integrated Services, IntServ)和区分服务(Differentiated Services, DiffServ)。Web QoS是随着QoS控制技术诞生的一种面向Web客户和HTTP请求提供性能保证及服务区分的技术。

3.2 Web QoS面临的困难

由于HTTP请求呈指数性增长，Internet上的许多热门站点都经常面临着服务器超载问题。研究表明，人们期望的Web站点的理想响应时间为1秒，普通的Web用户通常不会忍受超过8～10秒的等待时间。具体而言，Web服务请求的响应时间主要由两个因素决定：网络传输的质量和Web服务器的处理性能。如果Web服务器不支持任何QoS控制，那么，在服务器过载的情况下，具备端到端网络QoS保证的高级流仍有可能遭受服务拒绝，或者Web服务的平均响应时间比用户的期望值高出多个数量级，从而导致事实上的"拒绝服务"效果。由于服务器的超载问题，Web服务器已经在某种程度上成为实现端到端QoS的瓶颈。与传统的TCP/IP和HTTP服务的平均主义哲学不同，电子商务应用通常要求对用户或服务进行区分优先级别的处理，这是因为所有的Web事务对客户或服务器而言不可能都同等重要[5]。

4 Web QoS 技术的应用

Web QoS 属于应用层的 QoS，它量度的是用户在与 Web 站点进行交互时所感受到的服务性能。因为网络基础设施的差异，影响 Web QoS 的因数很多，特别明显的是下载时间、交易时间（如银行结算、股票交易、网上购物等）、服务器的可用性、遇到的错误（如失败的连接、丢失的页面或组件、中断的链路、交易失败）等等[6]。从我校的数字化校园建设中，通过对 Web 日志分析和学校师生的流动特性，我们从 2 个方面提出了提升 Web QoS 的实践方案，如图 3。

其一是区分服务对象的机制和策略。依托电信运营商的网络，校园网学校通过北电网络 Contivity 5000 VPN 设备建立一套 VPN（Virtual Private NetWork）虚拟专用网络。通过建立一个穿过公用网络的安全、稳定的隧道来实现临时的、安全的快速点对点接入园区局域网，从而提高网络的传输质量和可靠性，降低网络时延和丢包率，从而为校外住户、分校区用户、出差远程办工用户、远程工作者以及新用户（客户、供应商和合作伙伴）提供一种直接、快速访问园区局域网 Web 服务的目的。

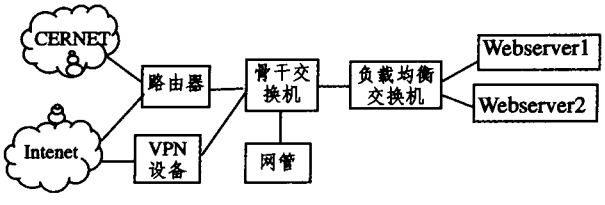

图 3　具有 VPN 和负载均衡功能的 Web QoS 示意图

其二是针对 Web 服务系统架构的设计机制和策略。就是在校园网教育数据中心里通过复制服务器内容技术建立 Web 服务器集群来增强本地处理能力，同时通过北电网络 Alteon Web SWitch 184 提供 Web 服务器负载均衡服务、QoS、内容加速等，实现了负载分担和带宽管理的功能，另外，Web 服务器集群内部使用的 IP 地址采用 192.168.0.x，通过负载均衡器实现内外地址的转换也提升了 Web 服务系统的安全性。

结束语　Web 挖掘技术是数据挖掘技术和 WWW 应用的结合，Web 日志分析是 Web 挖掘技术的一个组成部分，Web QoS 技术是应用层的 QoS，基于 Web 日志分析的 Web QoS 的成功应用，在网站建设和电子商务中起到了重要作用。

目前，Web QoS 控制技术的研究已经越来越多地受到网络研究者和著名公司的重视，IBM 公司也推出了支持 Web QoS 控制机制的名为“WebSphere”的软件平台。SUN 公司推出了 JES（Java Enterprise System）软件平台并进行了部分开源，HP 公司已经推出了在 Web 服务器中支持 QoS 控制机制的名为"WebQoS"的服务质量软件，同时 Cisco 公司、Nortel 公司正在推出具有 7 层交换功能的内容交换机、链路均衡器和高速缓存服务器等。

参 考 文 献

1　CNNIC. 第 19 次中国互联网络发展状况统计报告. 2007-1-23
2　美国太阳微系统公司. Sun ONE Application Server 开发指南[M]. 师炜. 北京：机械工业出版社，2003. 2～25
3　韩晓莉，李秉智. 个性化 Web 推荐服务研究[J]. 计算机科学，2006(2)：135～138
4　姜传菊. 网络日志分析在网络安全中的作用[J]. 现代图书情报技术，2004(12)：58～59
5　刘建国. Web 挖掘在电子商务中的应用[J]. 重庆工商大学学报，2004(4)：384～387
6　林闯. 服务质量(QoS)走进 Web. 计算机教育，2004(1)：38～40

（上接第 46 页）

图 1，图 2 分别为两种缓存算法在不同网络规模下 P2P 流媒体分发过程中节点获取流媒体前缀片段的平均传输时延以及所有片段的平均传输代价比较。最小化网络传输代价缓存算法在不同网络规模上无论是起动时延还是片段平均传输代价上都比自适应按需缓存算法小。由于网络规模不同，用户群聚度不同，流媒体片段局部流行度不同，因此此算法对流媒体分发性能改善程度也不同。从图 1，2 可知最小化网络总体传输代价启发式缓存算法能有效地减少流媒体在无线 Ad Hoc 网络中 P2P 模式分发的传输代价。

结论及研究展望　最小化网络总体传输代价启发式缓存算法与以往的无线网络流媒体分发缓存算法比较具有以下特点：1）以总体缓存资源有限为前提；2）结合考虑流媒体文件内部流行度，；3）考虑无线 Ad Hoc 网络中节点传输流媒体的可靠性；4）由于采用泛洪搜索算法，间接的统计了流媒体片段的缓存密度，其缓存原则就是缓存实际缓存密度与期望缓存密度之差和片段流行度乘积最大的片段，因此，此算法在总体缓存资源有限的情况下将使得流行度越高的片段传输时延越小，流行度越低的片段传输时延相对较长，这样就降低了流媒体分发过程中的总体传输时延。

本算法目前没有考虑无线 Ad Hoc 网络节点传输流媒体可靠性的探测算法，而是假定节点可靠性已知，且所有节点可靠性相同，但实际节点失效概率与无线网络中节点的密度和网络负载相关，随着时间和空间的变化而变化。因此预测节点失效概率将是我们下一步的研究方向之一。

本算法流媒体分发模型中，节点采用泛洪搜索算法寻找源节点，这样将给无线网络造成较大的通信负载，因此进一步研究有效的源节点搜索算法对减少无线网络通信负载，提高流媒体分发效率具有重要的意义。这也是我们下一步的研究方向之一。

参 考 文 献

1　Yang Z K, Wang T, Du X, Liu W, Yu J. Investigation on the content popularity distribution under K-Transformation in streaming applications. In: Proc. IEEE TENCON 2005, Melbourne, Australia, Nov. 2005. 1659～1663
2　PPLive. http://www.pplive.com
3　TVAnts. http://www.tvants.com
4　GridCast. http://www.gridcast.cn/index.jsp
5　Ghandeharizadeh S. Krishnamachari B. Shanshan Song, Placement of Continuous Media in Wireless Peer-to-Peer Networks, IEEE Transactions on Multimedia, APRIL 2004, 6(2): 335～342
6　Jin S. Replication of Partitioned Media Streams in Wireless Ad Hoc Networks. MM'04, New York, USA, Oct 2004. 396～399
7　Xiang Zhe, Zhang Qian, Zhu Wenwu, Zhang Zhensheng, Zhang Ya-Qin. Peer-to-Peer Based Multimedia Distribution Service. IEEE Transactions on Multimedia, APRIL 2004, 6(2)
8　Guo Lie, Chen Songqing, Zhang Xiaodong. Design and Evaluation of a Scalable and Reliable P2P Assisted Proxy for On-Demand Streaming Media Delivery. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(5): 669～682