CN43-1258/TP
ISSN 1007-130X
计算机工程与科学
COMPUTER ENGINEERING & SCIENCE
2006年第 28卷第 3期
Vol.28 No.3 2006

文章编号：1007-130X(2006)03-0026-02

Web网站流量分析系统设计*
Design of a Web Site Throughput Analysis System

蔡 巍，李名世
CAI Wei，LI Ming-shi
（厦门大学计算机系，福建 厦门 361005）
（Department of Computer Science，Xiamen University，Xiamen 361005，China）

摘 要：Web网站的流量分析系统是提高网站服务质量的可靠保证。目前，互联网上的大部份流量都是Web流量，而多数网站仅用简单的计数器来衡量其服务质量，这远远不能适应网站服务和投资评估的迫切需求。本文首先描述网站流量分析所应包含的指标体系，继而深入研究这一分析系统的具体实现技术，最后探讨了数据挖掘技术在分析系统中的可能应用。

Abstract：Web site throughput analysis system is a reliable guarantee for improving site service quality. At present，most throughput through the Internet is the Web throughput，and many sites estimate their service quality only by using a single counter. It is far from what is greatly needed in site services and investment estimation. This article begins with the criterion system of site throughput analysis，and then studies the specific implementation technique，and finally discusses the possibility of data mining in analysis systems.

关键词：网站管理；流量分析；网络监听；数据挖掘
Key words：site management；throughput analysis；network sniff；data mining
中图分类号：TP393 文献标识码：A

1 引言

在因特网发展初期，网站管理还处于起步阶段，这时还只有少数大公司和企业有自己的Web网站，管理者们用来统计网站流量的工具通常只是一个简单的计数器。现在，随着因特网深入到社会生活的各个层面，越来越多的单位和个人都建立了自己的网站。管理者们开始发现，分析网站流量是获得客户资料的一个有效渠道，可以从中得到各种有用的统计数据[1]。管理者可据此得知自己网站的受欢迎程度，以及哪些栏目办得最成功等基本信息；而公司网站的管理者更关心消费者们来自何方、自己的哪些产品最受欢迎等，从而改善公司的营销战略。在这种需求下，传统的那种简单依靠计数器来统计网站访问量的方法已经远远不能满足管理者们的要求，迫切需要一种能够统计分析各项信息的网站流量管理分析工具。

网站流量分析系统就是在这种趋势下应运而生的分析器，是为网站提供页面访问计数、排行和访问分析服务的系统，它可以对整个站点乃至任意页面的访问流量按需要进行各种技术分析，并向网站管理者提供完整的综合报告，使他们对自己网站的整体运作状况有一个清楚的认识。

通常，流量分析系统至少应该能够给出以下几方面的统计数据[2]：

（1）流量统计：包括网站的总访问量统计及其时间跨度；网站访问流量按时间段分组的排序，其中时间段可以选择以小时、天、周等单位。

（2）页面统计：包括网站每个目录的访问次数排序；网站每个文件的访问次数排序等。

（3）地区统计：包括访问者按源IP地址进行分类，并按访问频率进行排序；根据IP所处地区按省、国家分类，得出各个地区的总访问量。

（4）错误统计：包括各类错误代码统计，如404（文件没找到）、401（非授权访问）、500（内部服务错误）、403（禁止访问）等。

另外，还可根据管理者的特殊需要灵活选择其它额外信息的统计，如访问者使用的浏览器和操作系统的类型等。

* 收稿日期：2003-03-02；修订日期：2004-08-28
作者简介：蔡巍（1980-），男，福建建瓯人，硕士，研究方向为计算机网络管理。
通讯地址：350003 福建省福州市五四路158号环球广场33层福诺移动通信技术有限公司；Tel：13788878626；E-mail：caiwei1129@163.com
Address：33rd Fl，Global Square，158 Wusi Rd，Fuzhou，Fujian 350003，P.R.China

浏览者的访问信息中，有许多信息对网站管理者们都是很有价值的。例如，从源 IP地址可以推断出访问者大致来自哪个地区、省份或是国家；访问最频繁的页面暗示着访问者们最亲睐的信息、产品所在；按时间段排序的访问流量统计则可以推断出企业产品销售的淡旺季；而当发现错误代码404(文件找不到)出现比例过高时，则意味着网站上存在过多过期或错误的链接，该更新了。

通过对这些有效信息的统计，可以得到管理者们想知道的信息。

对于生成的统计报告，还可转换成各种图表的形式，以便管理者们可以更加清晰明了地知道网络流量的当前状况信息；按任意时段可以了解网站乃至任意页面的流量动向和受欢迎程度；访问者的来源根据源 IP追溯，对于国内可以精确到某个省市甚至更具体的地址，对于国外可以精确到来源于哪个国家。网站管理者尤其是企业销售网站的管理者，凭此可以知道哪些地区的人对哪些栏目感兴趣，从而有的放矢地做出广告投放等相应市场调整策略。

2 网站流量分析系统的实现模式

分析系统的实现大体可以分为两步：第一步是通过某种途径获得网站访问的相关记录，一条基本的记录包括访问者的源 IP地址、访问时间、请求的 URL等；第二步则运用现有手段、各种算法对网站记录进行详细的分析和统计。目前，流量分析的实现大体可以分成以下三种模式。

(1)通过监听网络数据包获得网站访问记录。这种方式第一步的实现是通过运行一个网络嗅探器类型的程序，监听所有经过的数据包，截获到包之后进行分析，根据服务器设定信息(如目的 IP地址、目的端口等)判断是否属于 HTTP信包，若符合，则进一步分析。一个典型的 HTTP请求如下[3]：

```
GET /somedir/somepage.html HTTP/1.1
Host: www.somedomain.com
Connection: close
User-agent: Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322)
Accept-language: zh-cn
```

从数据包可以提取出所需要的信息，如请求方式、请求资源、浏览器类型等，形成一个记录插入已构建好的数据库中(数据包的源 IP地址可以在第一步 IP层数据分析中得到，而接收到数据包的时间可以作为访问时间)，若不符合要求则抛弃。第二步是按照网站管理者的要求，通过调用数据库的各种检索、排序等指令，对相关的记录文进行分析、归类和统计，得到管理者们所关心的信息。这两部分程序可以是一个整体，但分别工作，互不干扰。当然，要在这种模式下进行流量分析，分析程序必须运行在 Web服务器或是所有访问 Web服务器的流量都必须经过的一台主机上。

(2)通过分析 Web服务端程序自动生成的日志文件获得访问记录。目前使用最多的 Web服务端程序如微软的 IIS(Internet Information Server，简称 IIS)、Netscape的 Fast Track Server以及 Linux系统下的 Apache等，在默认安装时都会配置成自动生成服务日志文件的运行模式，这些日志文件通常都存储着我们需要的一些信息。例如，默认配置下 WinXP中 IIS的日志文件格式如下：

```
2004-01-02 13:01:51 172.16.10.119 172.16.10.142 80 GET /postinfo.htm - 404
```

此格式分别对应着访问时间、访问者 IP、服务器地址、端口、请求方式、请求资源和应答。

因此，流量分析的第一步就可以通过分析这些日志文件的格式和所需信息，从而获得网站访问记录；第二步就是设计合适、快速的算法，对这些记录进行检索、排序、统计，得到所需数据。

(3)通过给网站页面添加统计脚本程序获得访问记录。这种方式是利用脚本文件(JSP、ASP或 PHP)编程获得当前访问者的各种信息，如 ASP脚本的内置 request对象就存储了当前访问对象的各种信息，编写一个脚本，把获得的各种有效信息作为一个记录存入数据库，如 request.ServerVariables("Remote_Addr")给出了访问者的 IP地址，而 request.ServerVariables("Http_User_Agent")则给出访问者所用的浏览器类型等。给网站的每个文件添加一段这样的脚本，就可以获得整个网站流量的统计数据；然后可以直接用脚本程序访问数据库，进行数据分析统计，也可以另外编写程序实现对数据库的操作。

这三种获取模式各有利弊：

(1)从分析程序安装的灵活性上看，第一种模式的缺点在于程序安装灵活度较小，由于是采用监听网络数据流的方法，因此程序要求运行在外部网络和 Web服务器之间的某台主机或是 Web服务器本身，这样才能保证记录到所有网站的访问流量；第二种模式就灵活得多了，分析程序可以安装在任何一台主机上，只要给它提供 Web服务端日志即可；而第三种模式最为麻烦，需要给网站的所有文件添加脚本语句，但可以改进，把访问记录的脚本专门写成一个文件，只要在网站每个文件中添加一个引用该脚本文件的链接就可以了。

(2)从对网站性能的影响上看，第一种模式下的数据包记录程序运行在 Web服务器上或是服务器前端设备必然会影响到服务器的性能，如降低请求响应速度等，可以考虑通过端口镜像减少对服务器性能的降低；第二种模式由于分析程序与服务器运行完全独立的原因，根本不用考虑是否会降低服务器性能的问题；而第三种模式由于每个网站页面被访问时都要额外有一个生成记录插入数据库的过程，网站性能不可避免地要有所降低，但影响不会很大。

(3)从分析要求的灵活性考虑，第二种模式能够分析的信息只能是日志文件中所含有的信息，分析的范围比较狭小；第三种模式鉴于脚本程序功能强大，可以获得的信息量比第二种模式多一些，但也只限于脚本程序所能提供的信息；而第一种模式则给出了网站管理者们更多的选择，可以灵活地定制所需要分析的信息，从而决定从符合要求的数据包中抽取哪些内容，如分析日志无法得到的操作系统和浏览器类型、HTTP版本、请求参数等。特别地，当要获得一些只能从分析数据包得到而又相对重要的信息时，第一种模式是唯一的解决方案。例如，获得网站对资源请求的响应时间是分析网站性能的一个重要参数，若平均响应时

(下转第第 46页)

5 结束语

本文提出的共享缓冲区信元重组结构大大提高了缓冲区资源利用率及信元重组效率，不但可以应用于 ATM接口的设计和实现，而且还能改进支持报文重组的交换开关；固定的延时上限，有利于信元重组结构对 QoS的支持。

参考文献：

[1] SKeshav, R Sharma. Issues and Trends in Router Design[J]. IEEE Communication Magazine, 1998, 36(5): 144-151.

[2] 孙志刚，卢锡城. 路由器 IP报文的重组和调度[J]. 计算机工程，2002, 28(5): 38-40

（上接第 27页）

间过长则意味着网站性能有待提高，否则会失去许多耐心有限的客户。通过第一种模式分析流经的数据包由请求包到达时间与回应包发送时间之间的差值，可得相应资源的响应时间，这种分析是第二、三种模式没有办法进行的。

（4）从分析速度上考虑，第一和第三种模式下，记录由于是存储在特定的数据库中，对这些记录的分析统计可以直接使用数据库的标准 SQL命令，因此效率较高，花费时间少。第二种方式通常都不借助数据库，因此对从日志中提取的记录信息进行分析一般是采用自己设计的算法。一般来说，当数据量很大时，这种方式花费的时间要比第一种方式多，但可以通过设计合适、先进的算法来提高分析速度。

第二种模式对网站性能无影响、程序安装灵活，目前是应用最多的一种网站流量分析模式。至于分析信息受限的缺点，可以通过定制日志文件格式来弥补：如 IIS和 Apache都提供了用户自己定制服务日志的方式，网站管理者可以根据需要定制日志的格式，使其可以包含所需信息，当然，可定制格式必须是服务器程序所提供的。第一、三种模式由于对网站性能有影响——这是业务量巨大的商业网站所不希望的，所以应用不广。其中，第三种模式由于不需要额外的程序，只要在原有网站文件中添加一些脚本，在网站文件不多的情况下实现较方便，因此被一些规模较小的网站采用较多。

3 网站流量分析系统的深入探讨

Web服务是当前网络服务中应用最广泛的方式，几乎所有常见的网络服务都是基于 Web平台提供的。据统计，网络上 70%以上的流量都属于 Web应用数据流量，而目前的流量分析系统功能有限，能提供的信息还不能满足管理者的特殊要求，他们需要知道更为详细和全面的信息，对网站发展提供指导，例如访问者在网站各个页面的停留时间。毕竟访问频率只是粗略的统计，也许十个匆匆而过的浏览者还没有一个在您的网站上停留了半个小时的访问者成为客户的可能性大。目前仍没有成熟的技术可以记录这种信息；此外，前面提到的网站资源响应时间的数据记录也没有很好地融合进当前的流量分析系统中。

流量分析系统还有待于和数据挖掘技术结合起来[4]，在信息数据开采功能上得以大大增强，这对以网上销售为主的电子商务网站来说特别重要。数据挖掘技术能从获得的网站访问记录数据中发掘出更有价值的信息，为网站管理和经营者们提供有利于公司企业长远发展的销售策略。这些增强的功能包括：

（1）自动预测趋势和行为。数据挖掘自动在大型数据库中寻找预测性信息，它根据时间序列型数据，由历史和当前的数据去推测未来的数据。一个典型的利用数据挖掘进行预测的例子就是目标客户判定。数据挖掘工具可以根据过去网站访问记录中的大量数据找出其中最有可能成为未来潜在客户的访问者。目前，预测方法有经典的统计方法、神经网络和机器学习等。

（2）关联分析。它反映一个事件和其他事件之间的依赖或关联的知识。如果两项或多项属性之间存在关联，那么其中一项的属性值就可以依据其他属性值进行预测，也就是通过数据挖掘可找出数据库中隐藏的关联网从而指导决策制定。例如，在分析记录时发现定购摄像头的顾客中有 90%的人同时也定购了打印机，这样就可将摄像头、打印机这些顾客经常同时定购的商品放在同一个页面来提高售货效益。

（3）偏差检测。数据库中的数据常有些异常记录，从数据库中检测这些偏差很有意义。偏差包括很多潜在的知识，如分类中的反常实例、不满足规则的特例、观测结果与模型预测值的偏差、量值随时间的变化等。通过这些异常实例的分析，经营者可以更加实时、精确地了解顾客的心态变化，更加迅速地调整销售战略以适应新的需求。偏差检测的基本方法是寻找观测结果与参照值之间有意义的差别。

4 结束语

目前的 Web网站流量分析系统虽然已经发展成为一个相对成熟的体系，能够满足普通网站管理者的大部分要求，但对于有特殊要求的用户来说仍不够完善，有待于和上述讨论的各种其它技术结合起来，发展成为为企业管理者们发掘出更多有价值的信息、为公司的发展出谋划策的强大利器。

参考文献：

[1] 王通. 企业网站网络营销策略 1——网站流量分析[EB/OL]. http://www.web136.net/wlch/20039172125 09.htm, 2003-09.

[2] Barry Craft. 全方位网站流量分析[EB/OL]. http://www.marketingman.net/wmtheory/wsp111.htm, 2003-09

[3] 张海波. 网络流量检测与分析：[硕士学位论文][D]. 西北工业大学，2001.

[4] 罗运模. SQL Server 2000数据仓库应用与开发[M]. 北京：人民邮电出版社，2001.