基于HTTP的网络数据包还原技术研究

谭敏生，汤 亮
（南华大学 计算机科学与技术学院，湖南 衡阳 421001）

摘 要：Web服务是Internet最常用的服务之一。针对用户Web行为的监视问题，阐述了信息编码的基本原理、数据包捕获与重组的基本方法，重点设计并用Java实现了GB2312，BIG5和UTF－8等编码方式的HTTP通讯信息的还原算法。结果表明文中研究的网络数据包还原技术能够正确还原HTTP通讯信息，达到了对Web行为进行监视的目的。

关键词：HTTP；信息编码；数据包；还原
中图分类号：TP393 **文献标识码**：A **文章编号**：1673－629X(2007)06－0176－03

Research of Network Data Packet Revert Based on HTTP

TAN Min-sheng, TANG Liang
(School of Computer Science and Technology in Nanhua University, Hengyang 421001, China)

Abstract: Web service is one of the most important services of the Internet. Aimed at monitoring the Web behavior, introduces the basic coding principle of information, the method of network data packet capturing and regrouping. It mainly designs and implements the revert of GB2312, BIG5 and UTF－8 coded HTTP communication by Java. The results show the HTTP communication could be reverted rightly and the Web behavior of network users could be monitored.

Key words: HTTP; information coding; data packet; revert

0 引 言

据国外数据统计，在开通互联网办公的企业中，企业员工平均每天有超过二分之一的上班时间用来上网聊天，浏览娱乐、赌博等网站，处理个人事务，员工用于下载与工作有关的文章与资料的时间只占下载时间的25%[1]。我国各行业的办公网络也存在着类似的情况。浏览新闻、搜索引擎、收发邮件、即时通讯、论坛/BBS/讨论组，是我国网民经常使用的五大网络服务/功能[2]，其中有三项是基于HTTP的Web服务。随着Web技术的迅速发展，一种越来越强的趋势表明，在不久的将来，Web有可能取代各种不同的服务器端和客户端的软、硬件平台及相应的应用系统，成为人们在Internet上进行信息发布和获取的标准平台[3]。因此，实时监控所在网络的HTTP通讯，当网络上出现非法内容的HTTP通讯时，将捕获的非法内容保存到数据库，向网络安全管理部门报告，以便及时保护企事业单位及国家的利益，已显得十分有必要。

收稿日期：2006－08－29
基金项目：国家自然科学基金(60572137)；湖南省教育厅科学研究项目(05C487)；衡阳市科技计划项目(2005KG01－015)
作者简介：谭敏生(1965－)，男，湖南衡阳人，教授，硕士生导师，研究方向为计算机网络、信息安全。

1 信息编码基本原理

计算机内部只能处理二进制数据，因此数值、文字、图像等数据，只有通过相应的编码标准，转换成二进制数据后计算机才能进行处理。常用的英文信息编码一般采用ASCALL，而中文信息编码一般采用GB2312(简体中文)、BIG5(繁体中文)、UTF－8(Unicode)等。

ASCALL又称ASCII码，为目前各计算机系统中使用最广泛的英文标准码，包含了英文的大小写字母、数字、标点符号等常用的字符，使用7位二进制数来表示。大小写英文字母的编码范围分别为41H～5AH和61H～7FH(H表示十六进制数，下同)。

GB2312又称国标码，通行于中国大陆与新加坡，它是一个简化字的编码规范，包括特殊符号、汉语拼音字母、俄文字母、日文假名等。GB2312规定对任意一个图形字符都采用两个字节表示，每个字节均采用七位编码表示。GB2312代码表分为94个区(A1 H～FE H)，每个区94个位(A1 H～FE H)。汉字的编码范围为：高字节B0 H～F7 H，低字节A0 H～FE H。

BIG5又称大五码，主要为香港与台湾地区使用，它是一个繁体字的编码规范，包括标准字、特殊符号、俄文字母、日文假名、使用者自造字等。BIG5编码由

两个字节构成,高字节的范围为 81 H～FE H,低字节的范围不连续,分别为 40 H～7E H 和 A1 H～FE H。汉字分为常用字和次常用字,使用范围分别为 A440 H～C67E H 和 C940 H～F9D5 H。

UTF－8 是 Unicode 的一种变长字符编码,理论上最多可以到 6 个字节,它可以处理地球上任何特殊的语言文字和符号公式。Unicode 使用一个字节的编码来表示 ASCII 字符,故 UTF－8 兼容 ASCII。当 UTF－8 编码长度超过 2 字节,则每个字节由一个换码序列开始,在首字节中,换码序列为 n 位 1 加 1 位 0(n 位 1 表示字符编码所需的字节数),后续字节的换码序列均为 10,各字节的剩余位是自由位。汉字的编码采用 3 个字节编码。

2　数据包的捕获与重组

文中使用 WinPcap 库完成网络数据包捕获的工作。WinPcap 是 Windows 平台下一个免费使用的函数库,采用分组捕获机制,通过访问网络的数据链路层来实现数据包的捕获[4]。它有以下功能:捕获原始数据包,包括在网络上各主机发送/接收的数据包;在数据包发往应用程序之前,按照自定义的规则将某些特殊的数据包过滤掉;收集网络通迅过程中的统计信息。WinPcap 结构简单,提供了一套与硬件无关的独立于系统的 API 封装函数,利用这些 API 函数即可完成所需的网络数据包监听功能。

有时一个有意义的 HTTP 通讯数据可能分布在几个数据包中,当 PDU 大小超过子网限制时,原始数据包将被分割成若干个小的数据包,每个数据包中含有自己的序号和下一个数据包的序号。由于 Internet 是基于分组交换的,数据包到达信宿机的先后顺序与序号之间没有直接关系,当数据包传到信宿机后,表现为一个无序的数据流。为了得到一个有意义的 HTTP 通讯数据,需要将数据包进行重组,可以设置一个缓冲队列,该缓冲队列的最大空间可设为滑动窗口的最大值。当接收到一个数据包时,比较其序号和应获得的序号,假如序号相同,则将其归入已排序的数据包行列,并从缓冲队列中将满足出队条件的数据包出队,若序号不同,则将其纳入缓冲队列中,并按序列号顺序排序,判断还需要哪些数据包[5,6]。

此外,在网络监视中,数据的发送者和接收者都是第三方,而监视可能是任何时候发起的,也许在对某一主机进行监视之前,其数据传输早已开始,或者在对某一主机进行监视之后,其数据传输还迟迟不开始,因而必须判定数据的起点和终点并抛弃掉不完整的应用数据。对于 HTTP 的数据内容,一种为请求数据(Request),一种为响应数据(Response)。对于请求数据,数据内容以“GET”,“POST”,“HEAD”,“HTTP”开头的即为起始数据包。对于请求信息结束的判定方法有两种情况:若请求信息中含有 Content－Length 域,则可根据其值依次取出规定长度的内容,即可确定结束数据包;若请求信息中不含有 Content－Length 域,则可以用两个 CRLF 作为结束标志;对于响应数据,数据内容为“HTTP”的即为起始数据包。对于响应信息结束的判定方法同样也有两种情况:若响应信息中含有 Content－Length 域,则可根据其值依次取出规定长度的内容,即可确定结束数据包;若响应信息中不含有 Content－Length 域,则可根据该数据包是否是 FIN 包来确定。

3　数据包的还原

重组得到的原始数据包为二进制数据,为了便于存储,转换为十六进制数据。由于数据包中可能包含了文字、图片信息等,采用 ASCALL 解码函数对原始数据包进行初始化,观察“Content－Type”、“charset”等特殊字段,判断该数据包是否传送的是文本信息,采用什么编码方式,进而解码还原。

3.1　GB2312 编码的还原

GB2312 编码还原的基本思想如图 1 所示,首先在转换后得到的十六进制数据中取一个字节,如果该字节的十六进制值不在 A1 H～FE H 的范围内,则判定编码对应的是 ASCII 字符,调用 ASCALL 解码函数,得到对应的 ASCII 字符;如果该字节的十六进制值在 A1 H～FE H 范围内,则可以判定该编码是属于某个汉字的编码,然后取后续的一个字节(汉字在 GB2312 编码中是两个字节),根据这两个字节的值,调用 GB2312 解码函数,从操作系统的字库中得出对应的汉字。继续根据以上方法把所有编码转化为相应的中英

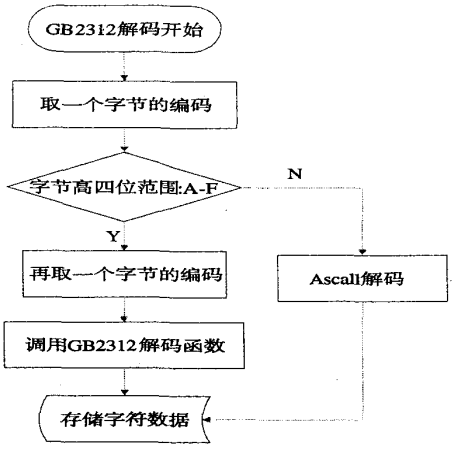

图 1　GB2312 编码还原算法流程

文字符信息。还原结果如图 2 所示。

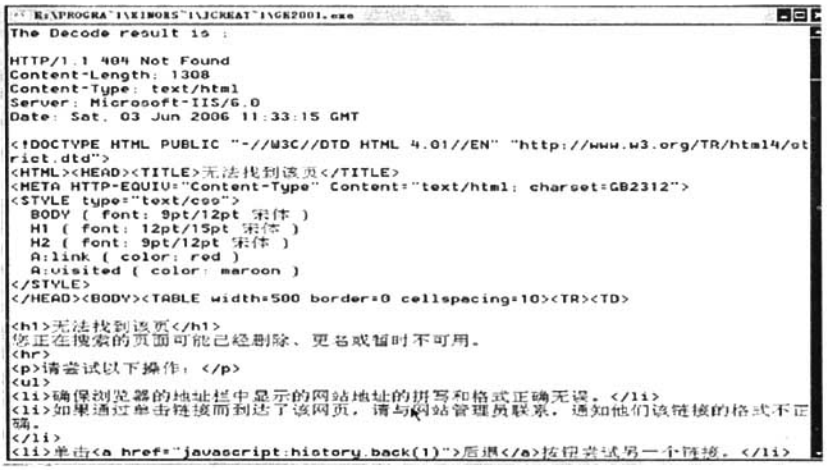

图 2 GB2312 编码的还原结果

3.2 BIG5 编码的还原

BIG5 编码还原的基本思想如图 3 所示，首先在转换后得到的十六进制数据中取两个字节，对取出的十六进制数进行判断，如果第一个十六进制数在 8 H～F H 范围内，第二个在 1 H～E H 范围内，第三个在 4 H～7 H 或是 A H～F H 范围内，第四个在 0 H～E H 范围内，则可判定该编码对应的是汉字，采用 BIG5 解码函数从操作系统的字库得出对应的汉字。如果编码不在汉字编码的范围内则视为 ASCII 字符，采用 ASCALL 解码函数解码，然后继续把所有十六进制数转化为对应的中英文字符信息。还原结果如图 4 所示。

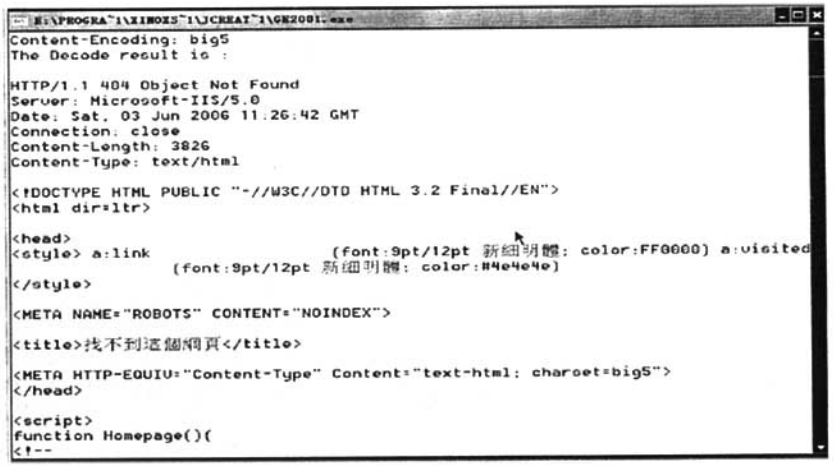

图 4 BIG5 编码的还原结果

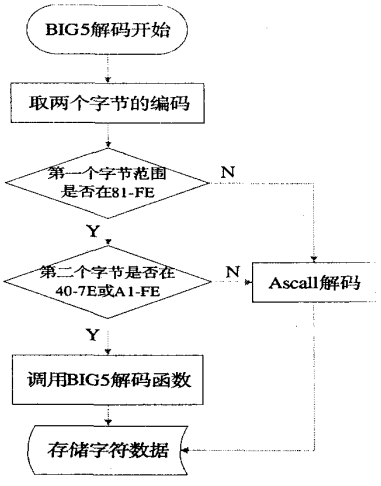

图 3 BIG5 编码还原算法流程

3.3 UTF－8 编码的还原

在进行 UTF－8 编码的还原之前，根据 UFT－8 编码转换表将十六进制数据重新转换为二进制数据。UTF－8 编码还原的基本思想如图 5 所示，首先从二进制数据中取一个字节，如果该二进制数不是以“1110”开始，就调用 ASCALL 解码函数解码为 ASCII 字符；如果该二进制数是以“1110”开始，那么可以判定这个字节是某一汉字编码的第一个字节，因为汉字采用三个

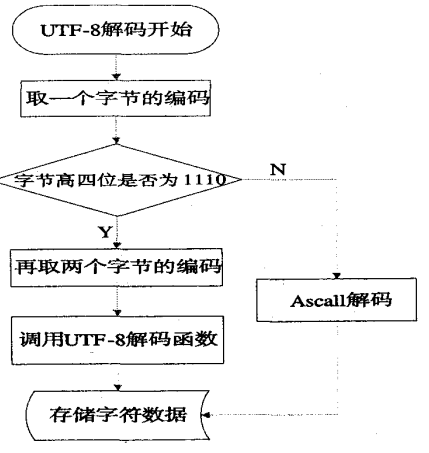

图 5 UTF－8 编码还原算法流程

字节编码，所以将后续的两个字节一起传递到 UTF－8 解码函数里，在操作系统字库中找到对应的汉字。还原结果如图 6 所示。

（下转第 231 页）

0.07 0.09)。

$$\begin{pmatrix} 0.13 & 0.28 & 0.28 & 0.21 & 0.10 \\ 0.13 & 0.32 & 0.29 & 0.18 & 0.08 \\ 0.10 & 0.24 & 0.26 & 0.26 & 0.14 \\ 0.14 & 0.27 & 0.27 & 0.22 & 0.10 \\ 0.14 & 0.28 & 0.30 & 0.24 & 0.04 \\ 0.03 & 0.27 & 0.32 & 0.21 & 0.17 \\ 0.10 & 0.33 & 0.33 & 0.18 & 0.06 \\ 0.05 & 0.27 & 0.28 & 0.27 & 0.13 \end{pmatrix} =$$

(0.13 0.23 0.23 0.23 0.17)

3.3 计算警务绩效评估结果

$CSD = B \cdot C^{T} = (0.13\ 0.23\ 0.23\ 0.23\ 0.17) \cdot (1\ 2\ 3\ 4\ 5)^{T} = 3.05$

则该派出所的警务绩效评估指数是

$$CSI = \frac{CSD}{5} \times 100\% = 0.61$$

4 结　论

实例显示,该派出所警务绩效评估值为 0.61,这个结果是目前警务绩效比较真实的反映。目前我国警务绩效评估值普遍偏低,原因是多方面的,由于国家对公安部门的财政投入明显不足,造成目前警力不足,办案经费短缺,刑侦和经侦设备落后,先进的侦查技术手段用不上,加大了案件的侦破难度;同时公安机关的信息化建设也明显落后于其他行业,金盾工程目前正在建设中,面对日益增多的计算机犯罪案件,公安机关缺少计算机专业人才;另外少数公安人员执法思想不端正,在执法过程中不能做到"立警为公、执法为民",也是造成警务绩效低的原因之一。可以看出,模糊综合评价方法用于警务绩效评估是有效的和实用的,其研究成果对于加强公安队伍建设,提高警察的工作效能,端正公安民警的执法思想具有一定的指导意义。

参考文献:

[1] 王舒娜. 警察绩效管理论要[J]. 公安研究,2004(8):57-62.

[2] 袁晓鹏. 浅淡公安工作绩效评价体系[J]. 中国人民公安大学学报,2000(1):93-96.

[3] 彭伊霖,陈少强. 论基层警务管理模式的变革创新[J]. 中国人民公安大学学报,2004(4):151-156.

[4] 朱立言,张　强. 美国政府绩效评估的历史演变[J]. 湘潭大学学报:哲学社会科学版,2005,29(1):1-7.

[5] 王　靖,张金锁. 综合评价中确定权重向量的几种方法比较[J]. 河北工业大学学报,2001,30(2):52-57.

(上接第 178 页)

图 6　UTF-8 编码的还原结果

4 结束语

研究了文本信息的还原技术,重点设计并用 Java 实现了 GB2312,BIG5 和 UTF-8 等编码方式的 HTTP 通讯信息的还原算法,结果表明这些算法能够正确还原用户的 Web 行为,因此,文中的研究工作为监视用户的 Web 行为提供了基本的技术手段,有利于规范用户的上网行为,促进文明上网,构建和谐的网络环境。同时,下一步工作将重点研究图像、声音等信息的还原技术。

此外,利用网络数据包还原技术对相关网络进行监视时,应遵循有关法律和法规。

参考文献:

[1] 王津林,赵满胜. 网络监控系统在局域网的应用[J]. 信息安全与通信保密,2004(2):58-59.

[2] CNNIC. 第 17 次中国互联网络发展状况统计报告[R]. 北京:中国互联网络信息中心,2006:18-20.

[3] 孙华领,顾景文. NDIS 驱动程序研究和基于 NDIS 网络监测程序实现[J]. 微计算机信息,2004(1):104-105.

[4] WinPcap. The Windows Packet Capture Library[EB/OL]. 2006. http://www.winpcap.org/default.htm.

[5] 范建华. TCP/IP 详解卷 1:协议[M]. 北京:机械工业出版社,2000.

[6] RFC2616. Hypertext Transfer Protocol——HTTP/1.1[S/OL]. 1999. http://www.rfc-editor.org/rfc/rfc2616.txt.