

# 基于 HTTP 协议的论坛群发技术的研究

陈慧玲<sup>1</sup>, 帅立国<sup>2</sup>, 姜昌金<sup>1</sup>

(1. 东南大学 自动控制系, 江苏 南京 210096; 2. 东南大学 仪器科学系, 江苏 南京 210096)

**摘要:**随着互联网的普及、上网人数的剧烈膨胀, 网络营销已经成为各个企业作为广告宣传、企业推广的必要手段, 群发技术也逐渐发展成网络营销的良好渠道。现在基于网络营销性质的群发技术有很多, 如: 商业网站群发、论坛网站群发、邮件群发、网络终端群发等等。文中主要是针对论坛网站对群发技术进行了深入的研究, 并提出了基于 HTTP 协议实现群发技术的思想, 介绍了该群发技术的框架和优点。同时分析了利用 Apache Jakarta 推出的新的开发包 HttpClient 的具体设计和实现方法。最后在网络安全方面, 从攻击者的角度分析了研究该技术的意义。

**关键词:** HTTP; 论坛群发技术; Java; HttpClient; 网络营销

**中图分类号:** TP393

**文献标识码:** A

**文章编号:** 1673-629X(2007)03-0037-03

## Research of Multi-Send to Forums Based on HTTP Protocol

CHEN Hui-ling<sup>1</sup>, SHUAI Li-guo<sup>2</sup>, JIANG Chang-jin<sup>1</sup>

(1. Automatic Control Department, Southeast Univ., Nanjing 210096, China;

2. Instrument Science Department, Southeast Univ., Nanjing 210096, China)

**Abstract:** With the popularization of Internet and rapid expanding of the number of people on network, net-sell has become an effective measure taken by all of corporations to advertise and propagandize themselves, and multi-send technology, as a convenient method, plays an important role. Now there are many sorts of multi-send technology, such as multi-send to business Web, forums, mailbox, net terminal and so on. Mainly researches the technology to forums and puts forward an idea of implementing the technology based on HTTP protocol. Particularly introduced the design and implementation of this technology using HttpClient development package released by Apache Jakarta corporation. Finally, the significance of researching this technology in the application of net security is analyzed from the point of attacker.

**Key words:** HTTP; forum multi-send technology; Java; HttpClient; net-sell

## 0 引言

网络营销是个高效低成本的事情, 绝大多数企业在国家企业上网工程的倡导下, 已经申请了域名, 建设了自己的网站, 但这仅仅是网站建设的一个开始<sup>[1]</sup>。做好网站的推广, 简单快速地做好网络营销, 才能快速地增加网站的流量, 而这些都离不开群发技术的支持。目前网络信息推广主要有: 邮件群发技术、引擎登陆技术、论坛群发技术、供求信息群发技术、留言板群发技术、QQ 群发技术、信使群发技术、IP 群发技术等等。邮件群发技术主要是指针对某些特定的人群发送特定的广告邮件, 行销的目的明确, 这样可以使得网站流量

迅速增加, 但是持续的时间比较短; 引擎登陆技术主要是指向上万个搜索引擎接口快速地提交网站数据信息, 使得搜索引擎能够搜索到网站的信息; 论坛群发技术是指快速地把网站的信息发布到各个论坛网站上, 现在由于论坛网站客流量的增加使得论坛群发技术成了一个很好的行销渠道。很多事情都是喜忧参半, 群发技术是网络营销的有利工具, 对企业的网络宣传起着至关重要的作用, 但是它也变成了垃圾广告的元凶, 是恶意信息的来源, 从攻击者的角度对群发技术的研究对网络安全也有深远的意义。文中主要研究基于 HTTP 协议的论坛群发技术, 并详述了其具体的实现方法。

## 1 基于 HTTP 论坛群发技术的框架

### 1.1 HTTP 协议简介

HTTP(Hypertext Transfer Protocol)协议是一个应用层协议, 其轻巧通用, 是整个 Web 的基础。HTTP

收稿日期: 2006-07-10

**作者简介:** 陈慧玲(1979-), 女, 河南濮阳人, 硕士研究生, 研究方向为网络客户端群发技术和网络安全; 帅立国, 博士, 副教授, 研究方向为精密仪器及机械; 姜昌金, 硕士, 副教授, 研究方向为控制理论与控制工程。

协议具有如下的特点:采用 HTTP 客户(如 Web 浏览器)与 HTTP 服务器(如 Web 服务器)通信;是一种无连接协议;遵循请求/应答模式;使用 MIME 内容类型。

HTTP 工作过程分为以下几个阶段(如图 1 所示)<sup>[2,3]</sup>:

- 1) 客户端向服务器端建立连接;
- 2) 客户端向服务器端发送请求信息;
- 3) 服务器端往客户端返回响应;
- 4) 断开连接。

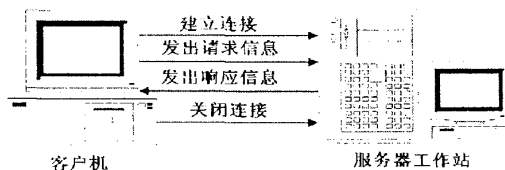


图 1 HTTP 连接模型

HTTP 协议的发展经历了三个阶段:HTTP/0.9, HTTP/1.0, HTTP/1.1。它早期的两个版本的特点是每次连接只处理一个请求,服务器处理完客户的请求,并收到客户的应答后即端口连接。采用这种方式可以节省传输时间。HTTP/1.1 对早期的两个版本有了很大的改进,即每次连接都是持久连接,可以处理多个请求,发送多个响应<sup>[2]</sup>。

在 HTTP/1.1 版本中,客户端向服务器端发送请求信息的方法有<sup>[2]</sup>GET, POST, PUT, OPTION 等,常用的是 GET 和 POST。简单的请求用 GET 方法,如果要往服务器发送表单数据,一般要用 POST。

## 1.2 技术框架及优点

各个论坛网站为了防止受到恶意的攻击一般都设置了用户权限认证,只有成为网站的会员才能在网上发帖,所以群发软件要设计一个注册机,可以自动地生成用户名和密码并进行注册,省去了每个网站依次去注册的麻烦;由于各个网站的注册和登陆的表单差异不是很大,可以用一个分析模块分析表单的格式达到注册和登陆的目的,但是对于差异比较大的网站可以编写定制的代码,对于发送信息的表单也是基于此。

群发技术的构建一般要具备以下几个模块:连接模块、注册模块、登陆模块、网站分析模块、发送模块。由于 HTTP 协议是 Internet 中通用的协议,它可以传输任何数据包括文本、图片、文件等。传输的文件内容比较丰富;基于 HTTP 协议易于穿透防火墙,因为任何网站要连到 Internet 上必须开通 80 端口。

## 2 群发技术的模型及实现

由于 Java 在网络方面具有强大的功能,因此文中

在实现方面采用了 Java 技术。客户端登陆、发送以及网站分析模块功能的实现主要采用了 HttpClient 技术。它是在 URLConnection<sup>[4]</sup>开发包的基础上对自动转向和 Cookie 技术进行了拓展,是专门设计来简化 HTTP 客户端与服务器进行各种通讯编程,支持客户端强大的丰富的功能。

### 2.1 群发技术的模型

整个系统由四部分组成:

(1)显示部分:该部分是一个人机交互的窗口,显示数据库中的有关论坛网站的信息,供用户选择要发布的网站,最后信息发布完后显示信息发布的状态。

(2)通信部分:这一部分内容较多,包括用户对数据库中保存的网站信息的调用,然后显示到交互界面上供用户查看选择;还有客户端和服务器之间的连接,由图 1 中的 Connection 模块完成,获得网站的页面信息保存到数据库中,并且会交给解析模块;然后是 Register 注册模块生成随机的用户名和密码进行注册,并把注册的信息保存到数据库中相应的网站表格里;接着是 Logon 登陆模块登陆网站并保存 Cookie 以便进行后面的操作;最后是 Postage 模块往各个网站提交发布内容。

(3)页面分析部分:该部分主要的功能是解析表单,把分析结果以一种结构化的形式存到数据库中,以供 Register, Logon, Postage 模块根据相应的表单格式提交相应的内容。

(4)数据库部分:数据库 1 存储了相关论坛的地址和主题内容,数据库 2 存储了连接的网站的有关信息和分析后的表单的结构信息。本系统采用了 SQL Server 2000,它是一个通用的关系数据库模型。模型如图 2 所示。

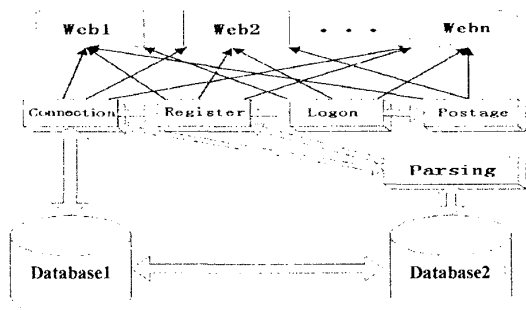


图 2 系统模型

### 2.2 群发技术的实现方法

本部分对各个模块的具体设计和实现方法作了详细的介绍。

#### 2.2.1 连接模块和分析模块

连接模块主要负责和服务器建立连接,然后定义

一个缓冲区读取网站的内容把网页发送给网页分析模块处理; Parsing 模块主要采用了 HTMLEditorKit.Parser 类来进行解析网页。它从 Reader 读入 HTML 文档在文档中寻找起始标记、结束标签、空元素标签、文本和注释,这五个部分已经涵盖了网页的大部分内容。连接模块的主要代码如下:

```
String url = "localhost"; //url 是定义的远程服务器的地址
HttpClient client = new HttpClient(); //客户端实例化
HttpMethod method = new GetMethod(url); //定义和服务器的交互的方法,这里采用 GET 方法
client.execute(method); ...
method.releaseConnection(); //关闭连接
.....
```

### 2.2.2 注册和登陆模块

注册机生成随机的定长(一般 8 位)的用户名和密码按照网站的表单格式进行注册,并保存注册信息到数据库中,成功后登陆模块进行登陆并采用 Cookie 技术保存用户的信息以便进行后面的提交,其主要代码如下:

```
PostMethod post = new PostMethod("/main.jsp"); //提交表单采用 POST 方法
NameValuePair name = new NameValuePair("name", "ld"); //定义表单的格式
NameValuePair pass = new NameValuePair("password", "ld"); ...
post.setRequestBody(new NameValuePair[] { name, pass, ... });
client.executeMethod(post);
.....
```

### 2.2.3 提交模块

提交模块按照网站的表单格式编写定制的代码,和登陆模块的格式一样。这里需要注意的是为了保存用户的 Cookie,登陆后不能再进行初始化,如果那样会丢失掉用户的登陆信息。

由于群发技术考虑到有大量的网站,所以本技术采用了多线程开发<sup>[5]</sup>, HttpClient 里提供了一个很好的类 MultiThreadedHttpConnectionManager<sup>[6]</sup>,该类对客户端定义了多线程管理。

在多线程技术中要注意共享资源的安全性和生命周期问题。因为不同的线程共享相同的内存,一个线程有可能会破坏另一个线程使用的变量和数据结构。一般来说,每个线程只有在确保资源不会改变或具有独占访问权的时候,才可以使用某个资源。同时也要防止死锁,就是两个线程都太小心,每个线程都等待对资源的独占访问权,却永远得不到。针对这种情况,文中利用了选择器(selector),一个线程可以查询一组 socket,找出哪一个已经准备好读写,然后按顺序处理就绪

的 socket。在这种情况下, I/O 必须使用通道和缓冲区,而不是流。

### 2.2.4 验证码的破解

许多网站在接受 HTTP 输入的时候常用到一项认证码技术,就是在显示提交数据页面之前或者同时,服务器会向客户端发送一个小图片,该图片通常由一些随机的数字或字母(很少情况下会有其他字符)组合而成,并要求在提交表单中正确输入该数字,无效的输入会使数据提交失败。要实现群发,就必须破解网站验证码。

早期的认证图片是以一个图片库的形式存放在服务器端,每次随机选取一幅图片发送到客户端,对于这种认证机制可以用自动机经过一定的穷举,下载所有的图片,并经过人工辨认存入数据库,客户端可以根据得到的图片甚至其校验和在数据库中查询其对应的认证码,发送相应的图片即可。现在的认证图片大部分由 jsp, php, cgi 等脚本动态生成的,即先随机产生认证码,再根据认证码动态生成图片。但是绝大部分网站,可能由于技术原因动态生成的图片在字体变形方面做的不是很理想。大多数甚至完全没有变形,这样识别这些认证图片就变得容易起来。

经过实际证明以某网站为例来具体说明破解方法:

图片格式有 bmp, jpg, gif 三种,由于格式之间可以互相转换, bmp 文件格式是最简单并且易于处理,这里只考虑 bmp 格式。笔者往该网站发送了 30 个请求,得到了 30 个图片,经过分析,图片内容为 4 位数字,不含字母或其他符号。图片格式为 bmp,所有图片具有统一的大小、宽、高。经过查询 bmp 文件格式文档,发现其调色板也完全相同,以 2 字节代表一个像素。另外还发现图片所有的数字所处的相对位置固定、大小固定、字体固定,变化的只是颜色和背景中一些随机的黑色斑点。而且基本上数字轮廓内的颜色以深色高灰度为主,变化并不大。背景中的斑点也是一样,数目约 40 个像素左右,进一步分析,确定各数字所处的位置,所占的像素。得到数据如下:

- (1) 图片高 15, 宽 51, 以下  $x, y$  是第一象限坐标;
- (2) 每个数字高 11, 宽 9, 字间间隙为 3;
- (3) 起始坐标依次在  $(x, y) = (4, 1), (16, 1), (28, 1), (40, 1)$ , 每个像素占用 2 个字节。

根据以上这些数据,可以定位每一个数字并提取出代表该位置数字的所有像素的数据信息,这些信息可以存在一个 198 字节的数组里。

那么如何准确判断这 198 个字节数据代表的数

(下转第 43 页)

要下拉至低于  $V_{Tn}$ , 则最大的上拉比不能超过 1.55。对于常规的工艺参数, 根据方程(7) 所得出的要求是容易达到的。可靠的写操作同样需要在迁移率之比、阈值电压和最高可达  $V_{DD}$  的最坏条件下进行模拟, 掩模偏差对上拉比的影响也应该考虑。所有最坏条件的考虑可能会得出苛刻的上拉比约束。

上面详细地分析了双口 RAM 存储单元的读写操作, 并以此为依据, 在充分仿真的基础上选择的管子参数如下:

$$\left(\frac{W}{L}\right)_{N1} = \left(\frac{W}{L}\right)_{N2} = \frac{1.5\mu\text{m}}{0.6\mu\text{m}}$$

$$\left(\frac{W}{L}\right)_{P1} = \left(\frac{W}{L}\right)_{P2} = \frac{0.9\mu\text{m}}{0.75\mu\text{m}}$$

$$\left(\frac{W}{L}\right)_{N3} = \left(\frac{W}{L}\right)_{N4} = \left(\frac{W}{L}\right)_{N5} = \left(\frac{W}{L}\right)_{N6} = \frac{0.9\mu\text{m}}{0.6\mu\text{m}}$$

### 3 结 论

异步 FIFO 缓冲存储器是现代系统设计中为提高整体性能而使用的一种重要手段<sup>[5]</sup>。文中分析了 FIFO 内部静态双口 RAM 基本存储单元的读写过程并给出了设计参数, 以提高 FIFO 的工作速度。实际工作中实现了一个  $1\text{k} \times 9\text{bit}$  的异步 FIFO (如图 7 所示), 版图设计采用了  $0.35\mu\text{m}$  CMOS 工艺规则, 芯片面积  $2041\mu\text{m} \times 2885\mu\text{m}$ 。仿真结果表明其读写速度约为

10ns。

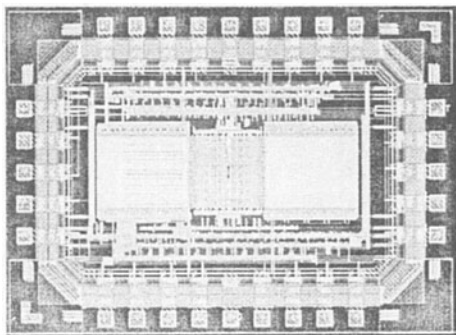


图 7  $1\text{k} \times 9\text{bit}$  的异步 FIFO 的实现版图

#### 参考文献:

- [1] 罗 昊. 一种异步 FIFO 的设计方法[J]. 电子技术应用, 2004, 30(8): 70-71.
- [2] 王传政, 董建民.  $256 \times 9$  位 FIFO 存储器的设计与研究[J]. 微处理机, 1996(1): 31-33.
- [3] Rabae J M. 数字集成电路——设计透视[M]. 第 2 版. 北京: 清华大学出版社, 2004.
- [4] Baker R J. CMOS 电路设计、布局与仿真(英文版)[M]. 北京: 机械工业出版社, 2003.
- [5] Kanopoulos N. A First-In, First-Out Memory for Signal Processing Applications[J]. IEEE Transactions on Circuits and Systems, 1986 CAS-33(5): 556-558.

(上接第 39 页)

字, 这涉及到一个模糊匹配的问题。在本例中, 匹配目标只有 10 个, 干扰元素影响也不是很大。经过多次试验, 发现产生 10 个标准的匹配目标就足够了, 这个匹配目标称之为数字模板, 共 10 个, 分别代表 0~9 数字的典型灰度分布, 每个数字模板也占用 198 个字节。这样以后将每次得到的 198 字节数据与之比较, 差异最小的即认为最佳匹配。

当然现在也有一部分网站上的验证码既有数字也有字符, 如果每次同一个字母的大小、倾斜角度没有变化, 都可以通过抽取特征值来判断。步骤就是: 把图片转换成灰度图, 然后根据明度图, 设定一个阈值转换成二值图。然后进行数次膨胀算法和缩小算法, 去掉杂边和杂点, 得到纯净的黑白数字图, 最后再进行字符识别就可以了。

### 3 结 语

文中主要对 Internet 客户端的群发技术作了深入的研究, 并基于 HTTP 协议设计并实现了群发技术, 有一定的商业意义。群发技术是一把双刃剑<sup>[1]</sup>, 它对

网络营销起着至关重要的作用, 有利于网站的推广, 但是如果恶意地发送大量垃圾广告的话却会起到相反的作用。从攻击者的角度对群发技术作进一步的研究, 也有利于网络的安全。

#### 参考文献:

- [1] 谢希仁. 计算机网络[M]. 北京: 电子工业出版社, 2002.
- [2] Hypertext Transfer Protocol - HTTP/1.1 [S/OL]. RFC 2068. 1997-01. <http://jakarta.apache.org/commons/httpclient/userguide.html>.
- [3] 黄景文. 基于 HTTP 协议和数据库的文件上传方法[J]. 广西科学院学报, 2005(3): 186-188.
- [4] Harold E R. Java 网络编程[M]. 朱涛江, 林 剑译. 北京: 中国电力出版社, 2005.
- [5] Malik D S, Nair P S. 数据结构 Java 版[M]. 杨 浩译. 北京: 清华大学出版社, 2004.
- [6] Apache Jakarta Common HttpClient [EB/OL]. 2004-03. <http://jakarta.apache.org/commons/httpclient/userguide.html>.