基于数字校园 URL 书签应用的设计与实现①

Application of URL Bookmarks in Digital Campus Network

康金辉 (陕西理工学院 网络信息中心 陕西 汉中 723003)

摘 要: 为了解决传统的网络书签应用依赖于特定的浏览器以及安全存取问题，根据校园网结构和运行特点，提出了 URL 书签应用的实现模型。通过在客户端守护进程中非混杂模式捕获数据包的方法解决了提取 URL 时浏览器的多样性问题，并通过分析 HTTP 协议规范提出了用户访问 Web 页时完整 URL 的提取算法。最后说明了通过用户注册实现 URL 书签 C/S 应用的存取方法。

关键词: 数字校园 书签 提取 HTTP 协议 非混杂模式

1 引言

在学校活动中，教师、学生利用校园网获取知识、开展教学活动已经成为学校教学、科研、管理不可或缺的一部分。在获取知识的过程中，校园网用户通常将自己需要的或者感兴趣的页面利用浏览器的收藏夹功能或者书签功能加以收藏。随着网络应用的进步，用户通常需要共享书签，异地利用书签、快照网页保存等。特别是共享书签为实现信息资源实现人性化的、更为广泛的、更为平等的共享和交流开辟了新的途径。网络书签又称网摘。网摘增加了一些 Web2．0 的典型特征如：社会协作、知识共享等特性。核心功能由原来的备份演变为发现信息、保存 URL 、快照网页、共享心得、聚合群体[1]等。常见的实现网络书签的控件有：和讯网摘控件、新浪 ViVi 网络收藏夹控件、365key 的天天网摘等。但是网络书签通常在本地存放，容易随着机器或浏览器的重装等而遭到破坏。在校园网环境下，需要一个面向全局用户的获取和管理网络书签的应用系统。

在校园网中是由网络用户端发起网络书签的管理申请，因此，如果要实现书签的网络应用，从原理上看，既可以是 C/S 模式、也可以是 B/S 模式。但实际应用表明，采用在客户端守护进程的 C/S 模式明显具有一定的优势。

2 系统结构模型

系统结构如图 1 所示。在图 1 中，虚线框内为客户端，整体系统具有 CA 服务器和书签服务器，是典型 C/S 模式的三层结构。在校园网环境下，通常书签服务需要 Web 服务器提供一个 Web 页面来说明书签服务。为了安装客户端，传统的方式是安装具有书签功能的经过 CA 认证的控件，但控件的作用仅仅限于所应用的有限的几种浏览器。例如：如果是仅仅针对 IE 浏览器的进程，当浏览器访问到需要的页面时，可以直接通过 API 函数捕获浏览器访问的地址，但用户的浏览器多种多样，如果是针对几乎所有浏览器并且通过 API 函数捕获的方法获得 URL 地址，不仅复杂，而且没有很好的扩展性，即无法适应用户浏览器的多样性。因此，鉴于浏览器支持 HTTP 协议的特点，如果要在校园网部署书签服务，本文提出，首先通过在校园网中部署 CA 服务实现 ActiveX 控件认证，当用户访问含有 ActiveX 控件的介绍 Web 页面时，首先在客户端自动安装 CA 服务的根证书，并且浏览器会从已具有代码签名的 ActiveX 控件中解读出其签名证书(公钥)和 Hash 表摘要，在默认的情况下，与 Windows 的“受信任的根证书”相比较查验公钥证书的有效性和合法性，验证签名证书正确后，就可以确认此代码确实是来自真实的软件开发商；之后，通过 ActiveX 控件下载位于远程 FTP 服务器上(也可以在 CA 服务器上)用于书签服务的应用程序，并且执行它，形成客户端的常驻进程。当然，也可以在书签服务的介绍页面由用户下载，安装执行形成守护进程。

当客户端的常驻进程工作时，可以采取捕获应用

① 收稿时间:2008-08-23

层数据包的方法通过解析获得 URL 书签。这种设计，使得用户浏览器只要支持 HTTP 协议规范，即能准确地解析用户所访问的 URL 地址。

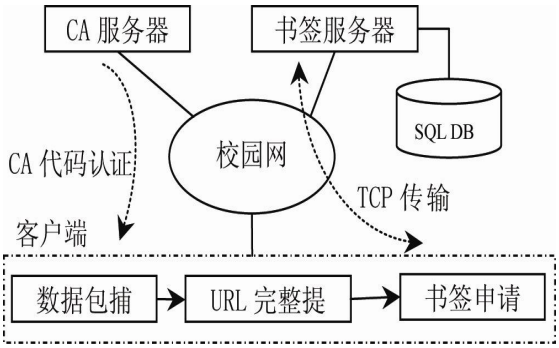

图 1　系统结构模型

捕获数据包的方式基于非混杂模式，这样，基于本地浏览器 Web 访问的数据包即被嗅探。而对于广播的非本地网卡流经的数据包流量则不予获取。特别要指出的是，由于在客户端运行了守护进程，可能对用户具有一定潜在的安全威胁，因此，基于 CA 认证的代码签名确保软件来自可信任的发行商，并且能保证软件代码不能被非法修改。

3　URL书签提取算法

HTTP 使用 TCP 协议的 80 端口进行可靠数据传输，一个 HTTP 会话由客户端开始发起，包括以下步骤：(1)客户端在浏览器中标识希望获取信息的 URL；(2)发起 HTTP 连接请求，启动客户端（UA）和一个初始 WWW 服务器或代理服务器之间的一个 HTTP 会话；(3)WWW 服务器或代理服务器根据客户端的 URL 请求将内容传送给客户端。

HTTP 协议定义了与服务器交互的最基本的方法是 GET 和 POST。事实上 GET 适用于多数请求，而保留 POST 仅用于更新站点。根据 HTTP 规范，GET 用于信息获取，而且应该是安全的。所谓安全的意味着该操作用于获取信息而非修改信息。由于守护进程运行于客户端，当浏览器利用 GET 方式或者 POST 方式提交请求时，捕获其数据包，一个典型的数据包结构如下：

GET /view/pics/11532/11532236.jpg
HTTP/1.1..Accept: */*..
Referer:
http://view.news.qq.com/a/20080713/000014.htm..
Accept-Language: zh-cn..
Accept-Encoding:gzip,deflate..User-Agent: Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)..
Host:img1.qq.com..Connection:Keep-Alive..
Cookie: pvid=367136496; flv=9.0; aduid=/F2WPdbT;

其结构遵从 HTTP 协议规范。在此不再解析。其中 Referer 头域允许客户端指定请求 URL 的源资源地址。也就是当用户访问某页面一个链接时，将当前页面的 URL 放在头域中提交给服务端。为了获得用户访问的 URL 及点击某个链接后访问某个页面的地址，其算法流程是：将捕获到内存中的应用层数据包经 Http 协议过滤后转换为字串，在头域中搜索“http://”的位置 P，从 P 位置开始，向整个字符串后搜索“Accept-Language”的位置 Q，如果搜索到则存贮从 P 到 Q 的所有字符串，否则退出处理过程。得到的 P 和 Q 之间的字符串即就是本次访问的完整的 URL 地址。这种方法可以获得客户端点击 Web 页面某个链接的信息，也就是用户点击一次，即能获得该用户点击该页面链接的完整的绝对 URL。

如：http://view.news.qq.com/a/20080713/000014.htm，获得 URL 后要考虑到头域中各个字段用“..”分隔。由于在客户机和服务器之间访问时建立了多个 HTTP 会话，因此，用户访问 URL 的一次请求中多个 HTTP 会话可能其 URL 相同，因此要实现过滤，过滤的结果是用户的一次 Web 访问中多个 HTTP 会话产生相同的 URL 地址过滤得到的结果只有一个。

过滤的方法是，要设计全局变量存贮上次 URL 文本字串，当 URL 文本消息被提取后，与上次 URL 比较，如果相同即退出过程，表明是一次 URL 访问的多个 HTTP 会话。如果不相同，则是用户新的链接访问。其流程如图 2 所示。

过滤的方法是，要设计全局变量存贮上次 URL 文本字串，当 URL 文本消息被提取后，与上次 URL 比较，如果相同即退出过程，表明是一次 URL 访问的多个 HTTP 会话。如果不相同，则是用户新的链接访问。

其流程如图 2 所示。

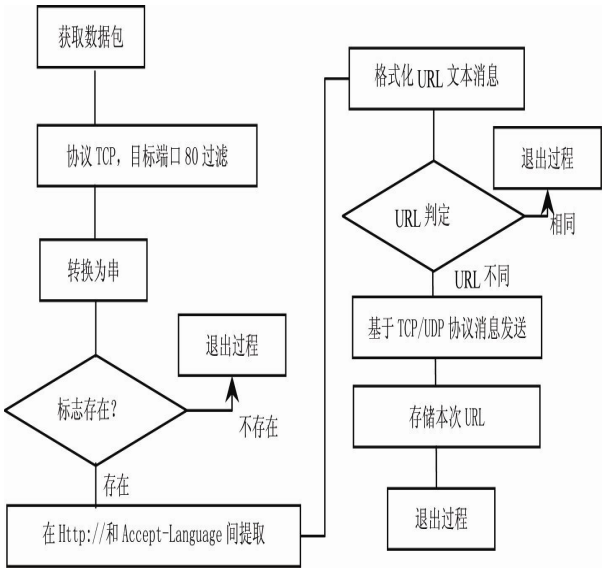

图 2　URL 提取流程

4　书签存取方法

由于用户点击的任何页面完整的 URL 被提取后存储记录，因此，如果用户想利用书签服务获得远程存贮管理，只需要点击浮动于任何浏览器窗口上方的守护进程运行窗口，打开最新获得的系统设定的 20 个 URL 地址窗口，申请即可。同样的道理，用户可以将远程存储于 URL 书签服务器的基于个人注册用户的 URL 书签下载到本地守护进程窗口中，进行修改、删除等操作。重要的是系统设计了 URL 书签的共享或者私有功能。如果用户的 URL 书签是共享的，则当用户随时随地存取远程的 URL 书签时，可以获得他人发布的共享 URL 书签。共享和私有的解决方法是通过网络用户的设定在服务器端用布尔类型的字段加以区分。

特别指出的是，服务器端后台数据库还设有非法 URL 站点的数据库表，其含有作者下载的大约 3 万个非法站点的 URL 书签，其作用是当网络用户共享含有这些非法站点 URL 书签时，通过服务器端内置的基于 UDP 协议的即时消息功能传递提示文本消息给客户端的守护进程，一方面在客户端自动打开 IE 浏览器将其重定向到一个校园网内部安全站点。另一方面在服务器予以屏蔽。

至于在客户端和服务器端通过 TCP 协议将获得的 URL 书签进行远程存贮管理，比较简单，不再赘述。

5　结束语

基于校园网 URL 书签的 C/S 方式在目前基于 1000M 链路的校园网环境中业已设计实现。采用了 Delphi7.0 编程工具基于 TCP 协议在服务器端能准确地实现书签的存取。虽然 TCP 协议是传输可靠的协议，但鉴于目前校园网优良的传输性能和带宽，作者也实验了基于 UDP 协议的数据存取，未发现数据包丢失的情况。但不可否认的是，用户端获得 URL 书签的守护进程应当配合校园网的单点登录系统、计费系统、即时消息系统等形成具有多种功能的客户端系统。虽然如此，系统也存在守护进程非法退出、防火墙隔离数据包、防病毒软件查杀、病毒、木马影响等问题。有待于进一步研究。

参考文献

1 沈阳.基于网络阅读行为兴趣度模型的网摘推荐.情报杂志, 2007,2:68－73.
2 郝志刚.互联网上社会书签的利用.中国西部科技, 2006,25:33－34.