基于Cookie的身份认证网站信息采集研究与实现

申 伟,李 翔,林 祥
(上海交通大学 信息安全工程学院,上海 200240)

摘 要:越来越多的需要认证才能访问的网站,对互联网内容安全管控平台中的网络信息采集单元提出了更高的要求。考虑到传统网络信息采集系统在应对身份认证网站时表现出的不足,基于Cookie内容协商机制,首创性地提出了面向身份认证网站发布信息的普适采集方案,并通过系列实验证明该信息采集方案的有效性与实用性。

关键词:内容协商;Cookie;信息采集
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2009)03-0178-04

Research and Realization for Information Collection of Website which Achieve Identity Authentication Based on Cookie Technology

SHEN Wei, LI Xiang, LIN Xiang
(School of Information Security Engineering, Shanghai Jiaotong University, Shanghai 200240, China)

Abstract: As the development of Internet technology, there are more and more Web pages that can't be visited without identity authentication. In order to implement content security management over these Web pages, the abilities of current information collection unit should be improved. In order to resolve the problem that traditional information collection unit can not retrieve Web page through identity authentication, proposes the project "the Information Collection Unit for the Website which Need Identity Authentication". After introducing of Cookie technology and Website's identity authentication negotiation, adds the Cookie content negotiation mechanism to the original information collection unit to retrieve Web page that the original unit can't support. In addition, provides an experiment to prove the effectivity and practicability of the new information collection unit.

Key words: content negotiation; Cookie; information collection

0 引 言

中国互联网络信息中心(CNNIC)2008年1月17日发布的《第21次中国互联网络发展状况统计报告》[1]显示,截至2007年12月我国网民人数达到了2.1亿,比上一年增加了7300万人,互联网已经飞速走进人们的日常生活,成为重要的大众传播媒体之一。

在享受信息时代巨大便利的同时,还要认识到互联网对于社会潜在的影响与威胁。为了在互联网上保持良好的舆论导向,维持社会的和谐发展,需要对互联网上的信息进行安全管控。互联网信息内容安全管控工作的重点是获取重要网络媒体的发布信息,并且基于知识发现技术在采集信息中实现互联网信息热点的自动挖掘。

收稿日期:2008-07-17
基金项目:国家自然科学基金项目(60502032);上海市科委项目(065115020)
作者简介:申 伟(1984-),男,山西运城人,硕士研究生,研究方向为互联网内容安全;李 翔,副教授,研究方向为网络内容安全。

1 研究对象

互联网信息内容安全管控工作主要包含三个方面的内容:信息采集、内容处理和结果呈现。信息采集为后续的内容处理和结果呈现提供最原始的数据,信息采集的好坏将直接影响到后续内容的处理。信息采集需要对大量的新闻网站和博客、论坛等交流网站的内容进行采集,以便能正确地反应互联网的舆论导向。

当前信息采集系统对网站内容进行信息采集的主要方式是采用网络爬虫软件从某一个网页开始,下载该网页内容及其网页内超链接指向的其他网页,递归下载直至对整个网站进行镜像[2]。这种方式对于一般的信息发布网站可以较好地进行信息采集,但对于多数需要身份认证网站遇到了一些问题。因为身份认证网站是需要用户登录之后才能查看到相关的页面,否则就跳转到该网站的登录页面。如果使用网络爬虫直接对身份认证网站发布信息进行采集,得到的往往是用户认证失败后跳转到的登录页面,而并不是想要获取的页面。

考虑到传统网络信息采集系统在应对身份认证网站时表现出的不足,基于Cookie内容协商机制,首创性地提出了面向身份认证网站发布信息采集的普适方案。

2 背景知识

作为文中重要的背景知识,Cookie技术[3]是随着Internet的Web服务发展而来的。Web服务主要通过HTTP[4]协议来实现浏览器和服务器之间的信息交互,由于HTTP协议是一种无状态协议,它并不能在同一用户不同时刻访问相同WEB网站时进行用户信息的记忆与继承,即HTTP协议无法对各个不同的用户进行区分[5]。这对于信息服务提供商和用户来说都极为不便,Cookie就是为了弥补HTTP协议的这一缺陷而诞生的。

2.1 Cookie的定义及工作原理

Cookie技术最早是由Netscape公司提出的,按照Netscape官方文档中的定义,Cookie是指在HTTP协议下,服务器或脚本可以维护客户端计算机上信息的一种方式。

Cookie机制主要通过报头来实现,定义了两种报头,分别是Set-Cookie报头和Cookie报头,它们位于HTTP协议数据包的头部。其中,Set-Cookie报头包含于Web服务器的响应头中,Cookie报头则包含在浏览器客户端请求头中,如图1所示。

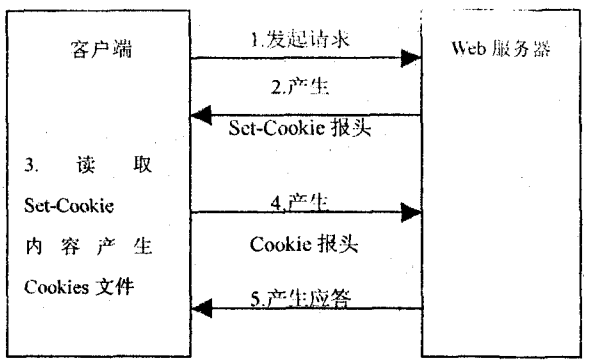


图1　Cookie机制的运行过程

(1)客户端发送获取网页的请求。

(2)服务器收到请求后,产生一个或几个Set-Cookie报头,放在HTTP报文中一起传回客户端。

(3)客户端收到应答后,将set-Cookie中的内容取出,形成一个Cookies文件存储在本地。

(4)当客户端再次向服务器发出请求时,先在本地寻找该网站的Cookies文件。若找到,则提取其中的Cookie信息,构成Cookie报头,放在HTTP请求报文中发给服务器。

(5)服务器接收到包含Cookie报头的请求后,查找该Cookie对应的用户信息,生成客户端所请求的页面传递给客户端。

2.2 基于Cookie实现身份认证原理

利用Cookie实现身份认证[6]是在传统Cookie机制基础上,Web服务器使用自行定义的认证方式,向通过认证的客户端传递一个代表客户端身份的Cookie信息。若客户端后续发送的请求中包含此Cookie,服务器就通过这个Cookie来识别特定的客户端,返回客户端请求的页面。若客户端后续发送的请求中不包含Cookie,或者包含了一个服务器无法识别的Cookie,则服务器认为这是一个新用户的请求,如果用户请求的是一个需要用户认证后才可以访问的页面,则服务器会返回一个认证失败的页面,并要求用户进行身份认证。如图2所示。

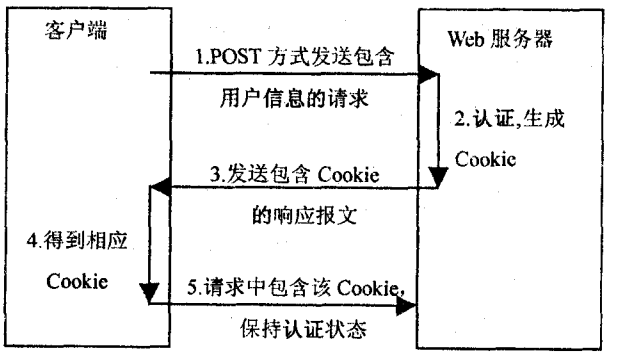


图2　基于Cookie认证的Web站点协商方式

3 基于认证Cookie实现身份认证网站的信息采集

对基于认证Cookie实现身份认证网站进行信息采集,关键是认证Cookie的获取,得到了认证Cookie后就可以使用传统的信息采集系统来获取需要认证后才可以访问的网站内容。考虑到Cookie信息是以纯文本的形式在客户端和服务器之间传送,因此可以使用网络侦听软件呈现客户端和服务器间交互的数据包,得到客户端和服务器间的信息协商方式,并从中截获所有Cookie信息加以分析,得到认证Cookie信息,用于身份认证网站的信息采集。

正如上文所述,在客户端成功登录后,服务器会返回给客户端一个登录成功的响应包,在这个响应包的HTTP头中包含一个或多个Set-Cookie报头,其中就包含有用于识别用户身份的认证Cookie。基于认证Cookie实现身份认证网站发布信息采集,所要做的工作是模拟客户端发送登录请求过程,在服务器响应报文中提取Cookie信息加以分析,寻找正确的认证Cookie信息。只要在登录请求报头中添加认证Cookie,就可以采集到需要身份认证才能得到的网站内容。

3.1 身份认证网站的信息采集过程

基于认证 Cookie 实现身份认证网站的信息采集主要包含模拟认证、认证 Cookie 分析和实际信息采集三个环节。其中模拟认证环节是通过编程模拟用户的登录过程,获取用户成功登录后服务器返回的响应包;认证 Cookie 分析环节是分析服务器的响应包中含有的 Cookie 字段,找到用于识别用户身份的认证 Cookie;实际信息采集环节是在后续信息采集请求报头中添加认证 Cookie,从而实现身份认证网站的信息采集。

3.1.1 模拟认证过程

(1)POST 包的监听。首先打开网络侦听软件侦听网卡数据包,然后在浏览器中登录需要进行信息采集的网站。找到浏览器发送给服务器端的 POST 包并保存下来。

(2)模拟发送登录请求 POST 包。编程构造和(1)中相同的 POST 包,使用 Socket 网络编程和服务器建立连接,把 POST 包发送给服务器,模拟用户认证的过程。

(3)接收服务器的响应包。在收到服务器发回来的响应报头中找到 set - Cookie 报头,读出其中的 Cookie 信息保存下来,并且转入认证 Cookie 分析阶段。

3.1.2 认证 Cookie 分析

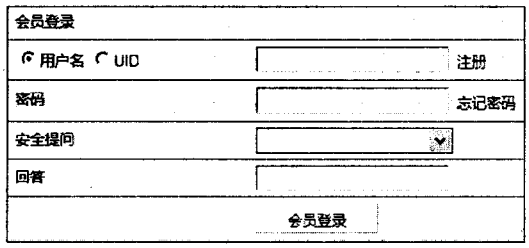
使用网络侦听软件分析用户成功登陆后服务器返回的响应信息中的 Set - Cookie 报头,一般情况下存在不止一个 Set - Cookie 报头,因为 Cookie 也经常用来记录用户曾经访问过的页面等信息。记下 Set - Cookie 中指定的 Cookie 的名称。再使用浏览器发送几个获取信息的请求,分析发送请求中的 Cookie 报头中包含的 Cookie 名称,一般 Cookie 值一直没有变化的 Cookie 名称就是进行信息采集所需要的用于识别用户身份的认证 Cookie。

3.1.3 实际信息采集

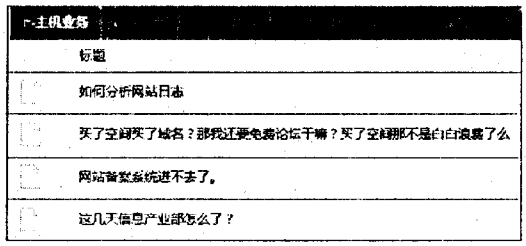
得到认证 Cookie 后,需要改造当前的信息采集系统,在发送的信息采集请求协议数据包中增加认证 Cookie 信息,模拟已经通过身份认证的用户来发送请求,实现需要身份认证网站的信息采集。

3.2 身份认证网站信息采集实例

文中选取 Crossday Discuz! Board 软件的官方论坛的"主机业务"频道(http://www.discuz.net/forum-46-1.html)作为需要身份认证网站信息采集的实验对象,该频道是需要用户经过认证之后才能查看的,若用户没有登录就会跳转到登录页面(如图 3(a)所示)。只有认证过程完成后,用户才可以浏览到需要登录才能查看的页面(如图 3(b)所示)。



(a)身份认证前



(b)身份认证后

图 3 身份认证前后同一页面的显示结果对比

3.2.1 模拟认证过程

首先使用网络侦听软件侦听浏览器发送的身份认证 POST 包(如图 4 所示)。

自行编程重构相同的 POST 请求包,基于 Socket 网络编程原理和 Web 服务器建立连接,将重构的 POST 包发送给 Web 服务器。

3.2.2 认证 Cookie 分析

服务器收到客户端的 POST 请求包后,验证用户身份的合法性,验证成功后发送给客户端响应的 HTTP 包,并在其报头中包含 Set - Cookie 头,给出实现身份认证后的用户认证 Cookie(如图 5 所示)。经分析,该响应数据包中含有 dznet-auth 的 Set - Cookie 字段就给出了用户认证 Cookie。

在发送编程重构的模拟认证 POST 请求包后,Web 服务器也会返回和上图相似的响应数据包。编程接受该响应数据包,并在其中提取认证 Cookie,即在响应包中查找"Set - Cookie"字段,找到后把 dznet-auth 及其所对应的 Cookie 值保存下来,供进行实际信息采集时使用。

3.2.3 实际信息采集

在现有信息获取请求协议数据包头部添加模拟认证过程中取得的认证 Cookie 报头,构造如图 6 所示信息采集请求数据包,实现需要认证的网站内容的采集。

在原有信息采集系统上添加认证 Cookie 后,采集到的页面不再是认证失败后 Web 服务器返回的登录界面,而是正常的"主机业务"版面。

3.3 结果分析

经过手动统计不难发现,需要身份认证才可浏览的 discuz! 论坛的"主机业务"频道(http://www.discuz.net/forum-46-1.html)首页包含 69 条信息发布

链接。在信息采集系统加载基于认证 Cookie 的身份认证网站发布信息采集机制前后,信息采集结果对比如表 1 所示。

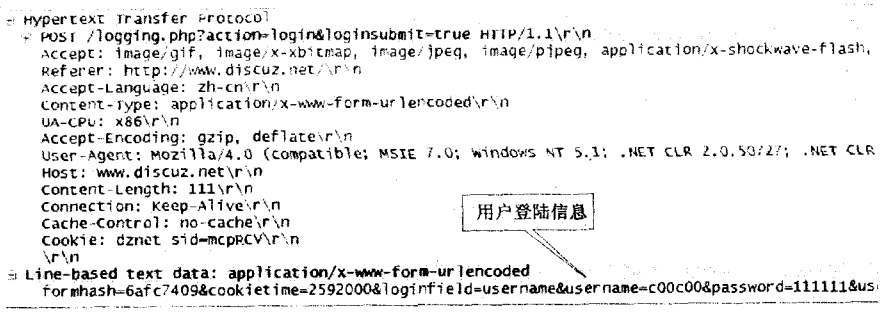


图 4　客户端发送的 POST 请求包

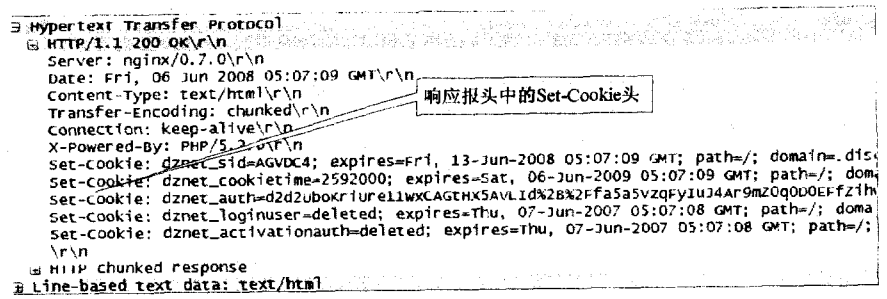


图 5　服务器验证用户身份后发给客户端的响应包

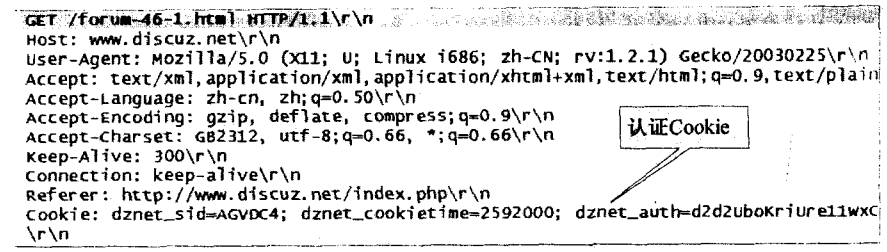


图 6　客户端发送的包含认证 Cookie 的请求包

表 1　加载基于认证 Cookie 的身份认证网站信息采集机制前后信息采集结果对比

信息采集	采集文件数	包含信息发布链接数	发布信息获取率
未加载身份认证网站信息采集机制	40	0	0%
加载身份认证网站信息采集机制	294	69	100%

由于没有进行身份认证,传统的信息采集系统只能采集到身份认证失败后 Web 服务器返回的登录页面及其之上的一些“注册用户”、“忘记密码”等几个无用的链接文件,无法得到有用的论坛帖子文件。

在传统的信息采集系统上加载基于认证 Cookie 的身份认证网站发布信息采集机制后,通过在发布信息请求数据包中添加认证 Cookie 字段,成功采集到需要身份认证才可浏览的信息发布链接全部内容。由于采集到实际需要的信息发布链接,信息获取环节得到的文件数明显增加,面向身份认证网站发布信息的链接获取率指数级提高。

综上所述,在传统的信息采集系统中加载全新的身份认证网站发布信息采集机制后,信息采集系统对于需要身份认证网站发布内容获取能力显著增强,有效改善了传统信息采集系统在处理需要用户认证才能访问的网站发布内容表现出的不足与缺陷。

4　结束语

就当前网络信息采集系统对于处理需要认证的网页所存在的局限性进行了讨论,在介绍了 Cookie 技术后,提出了基于认证 Cookie 实现身份认证网站内容的信息采集方案。该方案通过和服务器进行内容协商,模拟用户的认证过程,获得认证 Cookie,并在后续的页面请求中添加认证 Cookie,解决了传统信息采集系统不能采集需要认证后才可以访问的页面的问题,提高了网络信息采集的有效性。最后通过实验验证了基于认证 Cookie 实现身份认证网站的信息采集系统对于需要认证网页的采集,验证了该系统对原有信息采集系统的先进性,为互联网信息内容安全管控工作提供更有效的原始数据,促进互联网安全健康的发展。

参考文献:

[1] 中国互联网络信息中心. 第 21 次中国互联网络发展状况统计报告[DB/OL]. 2008. http://www.cnnic.net.cn/uploadfiles/doc/2008/1/17/104126.doc.

[2] 张　帆,李琳娜,杨炳儒. 基于 Web 的智能信息采集及处理系统设计与实现[J]. 计算机工程,2007,33(18):265-267.

[3] Kristol D, Montulli L. HTTP State Management Mechanism [S]. RFC2965,2000.

[4] Fielding R, Mogul J C, Frystyk H, et al. Hypertext Transfer Protocol - - HTTP/1.1[S]. RFC 2616. 1999.

[5] Thomas S A. HTTP Essentials[M]. [s. l.]: Wiley Publishing,2001.

[6] 谢丽春. 基于 Cookie 技术的用户认证[J]. 内江科技,2006(6):101-102.