

基于浏览器的用户身份识别系统

徐 晏¹, 张代远^{1,2,3}

- (1. 南京邮电大学 计算机学院, 江苏 南京 210003;
2. 江苏省无线传感网高技术研究重点实验室, 江苏 南京 210003;
3. 南京邮电大学 计算机技术研究所, 江苏 南京 210003)

摘 要:在现代社会生活中,互联网的运用空间不断拓宽,网络的安全也渐渐成为一个需要引起人们重视的问题。在诸多网络安全问题的防范中,用户的身份认证是一个基本的安全机制。传统的用户认证是基于用户名和与其对应的密码之类的方式确认用户身份,在此基础上,文中考虑对其他一些同样可以用于身份认证的因素加以利用,希望认证方式可以有多样性的,非单一的手段。文中研究的基于浏览器的用户身份识别系统就是通过收集浏览器的特征作为指纹对用户身份进行识别的,这就是一种新型的认证方式。

关键词:网络安全;身份认证;浏览器特征;指纹算法

中图分类号:TP31

文献标识码:A

文章编号:1673-629X(2013)08-0079-04

doi:10.3969/j.issn.1673-629X.2013.08.020

User Authentication System Based on Browser

XU Yan¹, ZHANG Dai-yuan^{1,2,3}

- (1. College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;
2. Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing 210003, China;
3. Institute of Computer Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: In modern social life, network security has gradually attract people's attention while the Internet plays an more and more important role in daily life. In many network security problems, the user authentication is a basic security mechanism. Traditional user authentication is based on the user name and their corresponding passwords or other similar ways to confirm the user identity. On this basis, also consider some other factors that can be used to make use of authentication, hoping can have a variety of authentication methods. For example, the research in this browser-based user identification system is worked by collecting the characteristics of browsers as the fingerprint, and this is a new authentication method.

Key words: network security; user authentication; browser features; fingerprint algorithm

1 概 述

由于当今计算机技术的迅猛发展,Internet的开放性以及其他方面因素导致了网络环境下的计算机系统存在很多安全问题。其中很多都是由于对用户的身份识别不足而引起的。

生物学上,人的指纹、声音、虹膜等具有唯一性、稳定性的特点,因而这些特点可以用于个人的身份识别。在计算机网络中,不同计算机的不同浏览器也有类似的特点。如果能对不同计算机的不同浏览器进行识别,就可以相应地识别访问网页的用户。

旧金山电子前沿基金会 EFF (Electronic Frontier Foundation) 新的研究发现,绝大多数的 Web 浏览器具有独特的特征。当使用浏览器浏览网页的时候,通过浏览器能够创建可识别用户身份的“指纹”。通过志愿者访问 EFF 的试验网站的结果发现,在近百万的匿名用户中,通过比较每个用户的浏览器版本,插件配置与用户的操作系统以及其相应的一系列内容,有 84% 的配置组合是独特的,可识别的。这也说明了在使用浏览器浏览网页时,浏览器的版本和结构信息可以被收集起来,设计为浏览器指纹进行用户的身份识别^[1]。

收稿日期:2012-10-24

修回日期:2013-01-30

网络出版时间:2013-04-22

基金项目:江苏高校优势学科建设工程资助项目(yx002001)

作者简介:徐 晏(1989-),男,江苏泰州人,硕士研究生,研究方向为神经网络、人工智能等;张代远,教授,博士,研究生导师,研究方向为人工智能、计算机体系结构、计算机应用等。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20130422.1722.036.html>

2 浏览器指纹

正常情况下,当用户通过浏览器来造访一个网站的时候,该网站便会在使用者电脑上留下 cookie 档案。并且将会以此记录用户的设定信息,并且追踪用户的线上行为,相应的,用户可以直接删除网站所遗留在电脑上的 cookie^[2,3]。尽管如此,浏览器指纹照样可以对其进行跟踪。同时浏览器指纹拥有隐秘的性质,这种性质使它可能比 supercookies 方法更难检测到,因为它没有在用户的计算机上留下持久的痕迹^[4]。

3 浏览器指纹的鉴定功效

可以设计一个关于浏览器指纹的分布算法,并且给予它足够的熵,这样就可以鉴别一个浏览器使用者的全球唯一性。这种全球性的鉴定可以被看作类似于一种 cookie,这种 cookie 不能被删除,除非对浏览器的配置进行了大量修改才可以打破这种指纹。

浏览器指纹可能会带来类似的“cookie 再生”的情况。这是由于在它的 IP 地址,子网地址甚至只是在它的自治系统编号的限定下,拥有 15 至 20 位信息的浏览器指纹足以唯一标识一个特定的浏览器。如果用户删除他们的 cookie,但是同时继续使用他们以前使用的 IP 地址,子网或 ASN。cookie-setter 很有可能把他们的新的 cookie 和旧的联系在一起^[5]。

指纹的最终用途是区分一个单独的 IP 地址背后的机器,即使这些机器完全阻塞 cookie。指纹很可能将为此工作,但也会有极少数的异常情况。

4 算法实现

4.1 浏览器指纹算法

在浏览器访问网站时,通过收集的一些常见的和不太常见的浏览器特性,可以实现一个浏览器指纹算法。本次课题考虑的浏览器特征包括:浏览器版本,浏览器插件,操作系统,屏幕分辨率与色深,用户时区,HTTP 接受类型头部(HTTP_ACCEPT_HEADER),系统字体,是否启用 cookie 等。把收集到的信息对应地划分为八个单独的字符串,这些字符串的一些包含多个相关的细节。该指纹本质上是这些字符串串联。

在某些情况下,该信息内容是简单的字符串,而在其他的测量情况下可以捕获更多细微的事实。例如,一个禁用 JavaScript 的浏览器将不记录有价值的视频,插件,字体和 supercookies 信息,因此这些测量的结果表明 JavaScript 是关闭的。再比如说,一个限制了 Flash 的浏览器在插件列表中显示有 Flash 插件,但却不能通过 Flash 获得系统字体的列表,从而创造出独特的指纹。同样的,当许多的测量和用户代理不相符,

就可以被识别出来这是伪造用户代理字符串的情况。

在浏览器指纹的收集过程中,还有很多其他的测量措施可以被考虑。有些特征我们没有察觉到它们可以被利用,或没有时间来正确的测量它——包括微软的 ActiveX 和 Silverlight API 方式收集的可指纹化措施(包括 CPU 类型和许多其他细节);在 Internet Explorer 中检测更多的插件;当 Flash 和 JavaScript 均不存在时由 CSS 自省完成系统字体的检测;浏览器发送的 HTTP 头的顺序;时钟偏差量的测量;TCP 堆栈指纹化等^[6]。作为一个实验,可以暂时不考虑这些,但是如果浏览器的指纹作为一项商业服务,则以上的各项因素应该全面考虑,也就是说,商业的浏览器指纹要比在这里研究的功能更加强大。

4.2 数学实现

通过设计一个算法可以将上一节所述的字符串按类型赋予一定的熵位,将字符串转化为数字信息^[7,8]。假设拥有一个浏览器指纹算法 $F()$,当有新的浏览器 x 计入计算, $F()$ 的结果是一个离散概率的密度函数 $P(f_n)$, $n \in [0, 1, \dots, N]$ 。下面给出公式详细计算出它的自信息量,或称作奇异量:

$$I(F(x) = f_n) = -\log_2(P(f_n)) \quad (1)$$

奇异量 I 是以比特为单位进行度量的,在上面的公式中通过以 2 为底的求对数得到。分布函数 $P(f_n)$ 的熵值是所有浏览器的奇异量的期望值,求值公式如下:

$$H(F) = -\sum_{n=0}^N P(f_n) \log_2(P(f_n)) \quad (2)$$

奇异量可以被看作是拥有指纹的对象身份的信息的总和,其中每一个比特位的确定可以将浏览器对应集合的可能性减少一半。如果一个网站被一组不同的浏览器 X 以相同的几率有规律的访问,直观地估计其中一个浏览器 x (x 属于 X) 将会以一种独一无二的身份被识别出来,如果 $I(F(x)) \geq \log_2 |X|$ 。应用二项式分布经过适当的时间间隔可以改变这种直观的认识,但是在实际的指纹应用中, $P(f_n)$ 的估测有着更大的不确定性。至少在尝试回答哪个浏览器是可以被独特的识别的时候情况是这样的。

当浏览器指纹是由不同的测量值组合而成的情况下, $F_s()$, $s \in S$,研究每个测量值的奇异值是有意义的,可相应地定义指纹各个组成部分的熵:

$$I_s(f_{n,s}) = -\log_2(P(f_{n,s})) \quad (3)$$

$$H_s(F_s) = -\sum_{n=0}^N P(f_{s,n}) \log_2(P(f_{s,n})) \quad (4)$$

要注意,当指纹的两个组成部分的奇异量 F_s 和 F_t 的测量是相互独立时,才能将这两个值线性相加。否则,就要用到条件自信息量。

$$I_{s+t}(f_{n,s}, f_{n,t}) = -\log_2(P(f_{n,s} | f_{n,t})) \tag{5}$$

比如鉴定是否安装了 Flash blocker 就可以先独立地进行插件测量和字体测量,然后再联合起来考虑。因为安装了 Flash 插件,而字体不能被检测到的情况是很小的,所以自信息量 P 可以被忽略不计。所以分别检测到这两者成立就说明 Flash blocker 存在。

5 软件实现与实验结果

本次课题操作系统采用 Windows XP。使用 Apache 搭建服务器^[9],以 module 方式将 PHP 与 Apache 结合使用,同时使用 JavaScript 进行脚本编辑。实验结果为一个网站,用户访问此站点,决定是否对自己浏览器进行分析,程序在得到用户授权后允许服务器列举、分析、测试 Web 浏览器的所有应用功能和插件情况,由于语言的局限性,对其特征的收集需要同时使用前台的 JavaScript 与后台的 PHP,在前台与后台同时收集完毕后将前台收集内容转入后台,然后将所有信息结合进行操作。对浏览器识别的信息收集之后,将收集的情况反馈给用户,同时搜索用户信息是否存在,即进行用户身份鉴定。最后将用户的相关信息记录下来,存入数据库。功能实现流程如图 1 所示。

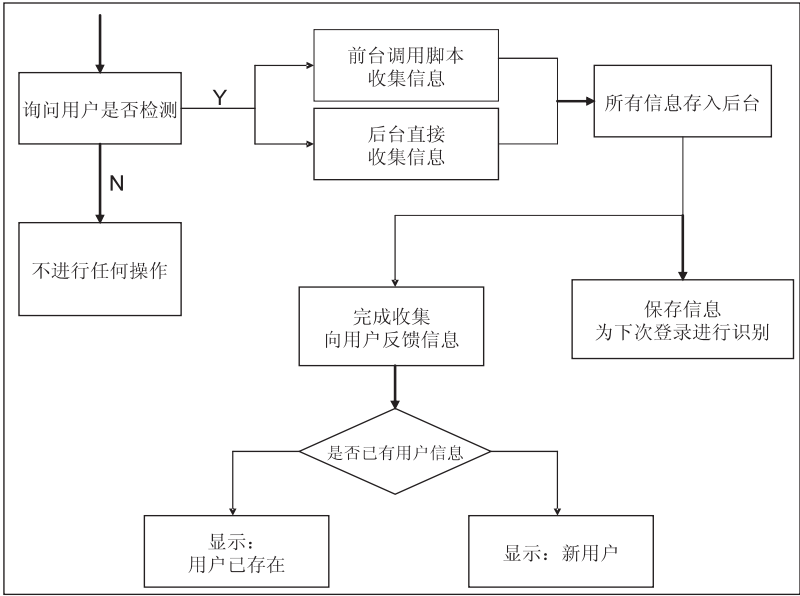


图 1 实现流程图

JavaScript 中的 screen 对象可以用于获取用户的屏幕信息。JavaScript 中的 Date 对象用于处理日期和时间。JavaScript 中的 Navigator 对象包含有关浏览器的信息。虽然没有应用于 Navigator 对象的公开标准,不过所有浏览器都支持该对象。检测 cookie 是否启用同样使用 Navigator 对象^[10]。

利用 PHP 所需要搜集的信息包括用户代理,HTTP_ACCEPT Headers,同时需要将前台信息导入。需要用到 \$_SERVER 变量。\$_SERVER 变量是一个特

殊的 PHP 保留变量,它包含了 Web 服务器提供的所有信息,被称为自动全局变量。\$_SERVER 是一个包含了诸如头信息 (header)、路径 (path) 以及脚本位置 (script locations) 等等信息的数组。这个数组中的项目由 Web 服务器创建。不能保证每个服务器都提供全部项目;服务器可能会忽略一些。对实验所需要的信息是通过 \$_SERVER 变量不同的参数设置后调用获得的^[11]。

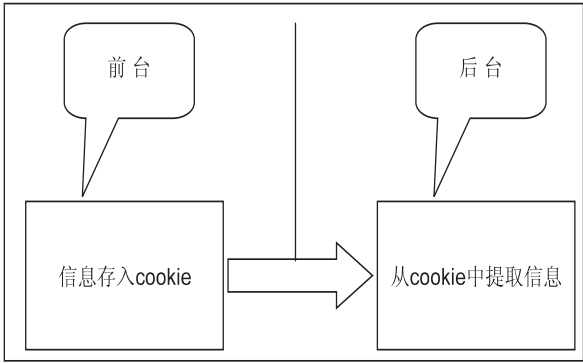


图 2 前后台交互图

前台与后台的信息交互方法有很多种,比如,表单传递、通过网址传递、cookie 传递、session 传递,考虑到本次课题需要交互的数据量并不大,所以采用 cookie 进行传输(见图 2)。这需要在前台的 JavaScript 和后台的 PHP 上都使用 cookie。因为需要 cookie 只保存当前访问者的信息,关闭浏览器之后最好就删除,所以完全可以不设置过期日期这一属性。

信息的收集与总和完成后按照其熵算法确定位数计算,将其保存到数据库之中,同时向用户反馈收集的信息。若用户信息存在,则提示“用户信息已存在”,然后更新时间信息;若信息不存在,则提示“新用户”然后将信息保存。

表 1 就是利用本机访问网站的显示内容。

6 结束语

文中研究浏览器的指纹,针对浏览器指纹特征讨论其应用。在此基础上设计一个算法,确定各项特征所需要的信息熵。继而设计一个用户身份识别的网站,利用网站的前台与后台共同工作,收集用户信息并加以处理,从而确定一个用户的身份。

实验结果展示了浏览器指纹确定用户身份的可行性与应用前景。

表 1 本机实验结果表

浏览器特征	具体值	鉴定位
浏览器与操作系统	Mozilla/5.0(Windows; U; Windows NT 5.1; en-US) AppleWebKit/534.16 (KHTML, like Gecko) Chrome/10.0.648.205 Safari/534.16	12.83
HTTP_ACCEPT Headers	text/html, */* GBK,utf-8;q=0.7,*;q=0.3 gzip, deflate, sdch zh-CN, zh; q=0.8	12.49
浏览器插件	Chrome PDF Viewer(pdf.dll),Google Gears0.5.33.0(gears.dll), Shockwave Flash(gcswf32.dll),Microsoft Windows Media Player Firefox Plugin(np-mswmp.dll),Ali Wang Wang Plug-In For Firefox and Netscape(npwangwang.dll),Alipay security control(npaliedit.dll),Thunder DapCtrl Plugin(npDapCtrlFirefox 2.0.5901.12.(744).dll),Silverlight Plug-In(npctrl.dll),Shockwave Flash(NPSWF32.dll),QQMusic(npQzoneMusic.dll),Default Plug-in(default_plugin)	21.1+
时区	-480	6.98
屏幕分辨率与色深	1280×800×32	4.72
是否启用 cookie	Yes	0.4
系统字体	Marlett, Arial, Arial CE, Arial CYR, Arial Greek, Arial TUR, Arial Baltic, Courier New, Courier New CE, Courier New CYR, Courier New Greek, Courier New TUR, Courier New Baltic, Lucida Console, Lucida Sans Unicode, Times New Roman, Times New Roman CE, Times New Roman CYR, Times New Roman Greek, Times New Roman TUR, Times New Roman Baltic, Wingdings, Symbol, Verdana, Arial Black, Comic Sans MS, Impact, Georgia, Franklin Gothic Medium, Palatino Linotype, Tahoma, Trebuchet MS, Webdings, Estrangelo Edessa, Gautami, Latha, Mangal, MV Boli, Raavi, Shruti, Tunga, Sylfaen, 仿宋体,黑体,楷体,宋体 & 新宋体, Microsoft Sans Serif, MS Mincho, MS PMincho, MS Gothic, MS PGothic, MS UI Gothic, Gulim, GulimChe, Dotum, DotumChe, Batang, BatangChe, Gungsuh, GungsuhChe, MingLiU, PMingLiU, Book Antiqua, Bookman Old Style, Century Gothic, Garamond, Haettenschweiler, Monotype Corsiva, Wingdings 2, Wingdings 3, LingoEs Unicode, Arial Narrow, Cambria, Cambria Math, Calibri, Candara, Consolas, Constantia, Corbel, Bookshelf Symbol 7, MS Reference Sans Serif, MS Reference Specialty, MT Extra, Euclid, Euclid Symbol, Euclid Extra, Euclid Fraktur, Euclid Math One, Euclid Math Two, Fences, Tiger, MT Extra Tiger, Symbol Tiger, Tiger Expert, Symbol Tiger Expert (via Flash)	21.1+
用户是否存在	该用户信息已存在	

参考文献：

[1] Eckersley P. How Unique Is Your Web Browser[EB/OL]. 2010. <https://panopticklick.eff.org/browser-uniqueness.pdf>.

[2] Mayer J R. "Any person... a pamphleteer" Internet Anonymity in the Age of Web 2.0[R]. Princeton:Princeton University,2009.

[3] 申伟,李翔,林翔.基于 Cookie 的身份认证网站信息采集研究与实现[J].计算机技术与发展,2009,19(3):178-181.

[4] Veridicom Inc. World Leader in Fingerprint Authentication Technology [DB/OL]. 2002 - 09. <http://www.veridicom.com/>.

[5] Zhao Qing. The Design of Security Authentication System Based on Campus Network [C]//Proc. of 2010 International Conference on Electrical and Control Engineering. [s. l.]: [s. n.],2010.

[6] Elsevier. Browsers uniquely identify users[J]. Network Security,2010,2010(5):19-19.

[7] 王美义,张凤鸣,刘智.模糊信息的熵权多属性决策方案评估方法[J].系统工程与电子技术,2006,28(10):1523-1525.

[8] 孙光辉.信息熵与不确定性[J].青岛大学学报:自然科学版,2000,13(3):50-51.

[9] 刘晓辉.网络服务搭建、配置与管理大全[M].Windows版.北京:电子工业出版社,2009.

[10] Powers S. JavaScript 核心技术[M].北京:机械工业出版社,2007.

[11] Lerdorf R, Tatro K, MacIntyre E. PHP 程序设计[M].第2版.北京:电子工业出版社,2009.

基于浏览器的用户身份识别系统

作者:

徐晏, 张代远, XU Yan, ZHANG Dai-yuan

作者单位:

徐晏, XU Yan(南京邮电大学 计算机学院, 江苏 南京, 210003), 张代远, ZHANG Dai-yuan(南京邮电大学 计算机学院, 江苏 南京 210003; 江苏省无线传感网高技术研究重点实验室, 江苏 南京 210003; 南京邮电大学 计算机技术研究所, 江苏 南京 210003)

刊名:

计算机技术与发展

ISTIC

英文刊名:

Computer Technology and Development

年, 卷(期):

2013 (8)

本文链接: http://d.wanfangdata.com.cn/Periodical_wjtz201308020.aspx