doi :10.3969/j.issn.1671-1122.2014.01.014

中小型网站智能安全检测研究

康海燕，祈鑫，魏美荣

（北京信息科技大学信息管理学院，北京 100192）

摘　要：随着互联网的发展及经济利益的驱动，黑客已将攻击重点转到 Web 应用服务器上，由此危害了服务器安全及客户端安全。针对这一现状，文章首先采用广度优先算法实现网络爬虫来获取目标网站的架构信息；然后用网页动态参数判定、网站架构分析、信息智能识别等技术对网站安全进行辅助检测，用正则表达式过滤非法跨站请求，实现跨站脚本攻击检测；最后，用正则表达式和 Python 强大的库资源编程实现了应用安全的实时检测和评估功能。实验表明：该系统在一定程度上减少了 Web 恶意攻击行为所带来的损失，提高了应对网页信息安全突发事件的响应速度。

关键词：篡改检测；XSS；网络爬虫；正则表达式

中图分类号：TP309　　**文献标识码：**A　　**文章编号：**1671-1122（2014）01-0061-04

Study on Small and Medium-sized Websites Intelligent Safety Inspection

KANG Hai-yan, QI Xin,WEI Mei-rong

(*School of Information Management, Beijing Information Science and Technology University, Beijing* 100192, *China*)

Abstract: With the development of the Internet and the economic benefits derived from it, hackers have been focused on in the Web application servers, which endanger the safety of the server and the client security and is against the status quo. First of all, the Web crawler works by using breadth-first algorithms to get the target site architecture information. Second, use page dynamic parameter determination, website structure analysis, information intelligent identification technology (such as auxiliary detection), and guard the security of the website with regular expressions to filter illegal cross-site requests. Then implement cross-site scripting attack detection. Finally, with regular expressions and powerful Python library resources programming create the real-time detection and assessment of the application security function. Experiments show that the system to a certain extent, reduces the loss on the Web due to malicious attacks and improves the response speed of the Web information security incidents.

Key words: tamper detection; XSS; crawler contrast; regular expressions

0 引言

目前，互联网主要面临的四大 Web 攻击类型分别为：网页篡改攻击、XSS 跨站脚本攻击、CSRF 跨站请求伪造攻击、电子商务网站钓鱼攻击。当上述攻击代码被嵌入 Web 通信中，传统的防火墙只是验证 HTTP 协议本身的合法性，无法判断对 HTTP 服务器的访问行为是否合法。此时，网络防火墙对 Web 应用起不到任何保护作用。由于 Web 服务器对用户请求的页面缺乏完整性保护机制，而防火墙、入侵检测系统以及入侵防御系统等网络安全设备对应用层面的攻击防范效果也并不理想，而且传统的解决方案存在着篡改检测不及时占用系统资源较大网站服务器造成负载较大、耗时多、危害大等不足之处。

面对严峻的网站安全形势，需要应用网页防篡改系统构建一个较为完善的网络安全体系，进一步解决传统两层防护体系存在的安全漏洞，从而提出我们的中小型网站智能安全检测系统。该系统运用正则表达式 [1] 和 Python 强大的库资源 [2]，在广度优先算法的基础上，设计线程池机制下的爬虫算法，实现对网页信息特征的爬取，并和自主设计的篡改规则匹配，

收稿日期：2013-11-25

基金项目：国家科技支撑计划课题 [2012BAH08B02]、教育部人文社会科学项目 [11YJC870011]、北京市教委科技计划面上项目 [KM201211232014]、校教学改革立项项目 [2012JGZD07]

作者简介：康海燕（1971-），男，河北，教授，博士，主要研究方向：信息系统安全和网络隐私保护；祁鑫（1991-），男，北京，本科，主要研究方向：信息安全；魏美荣（1992-），女，北京，教授，本科，主要研究方向：信息安全

实现对目标网页的篡改检测、敏感词检测、域名解析可用性、网站服务可用性和网站程序可用性检测等功能。广度优先算法满足了多级页面的结构需求，线程池属于对象池，最重要的特征是最大程度利用线程，使编程模型更清楚、更优化。该爬虫算法较之传统的篡改监测方法，有检测及时、耗时短、系统资源消耗少、功能更全面等特点。基于XSS、CSRF、钓鱼攻击[3]的原理和特征，该系统利用网页动态参数判定、网页结构分析、信息智能识别等技术实现对上述攻击的检测。该检测方法及时、准确，弥补了绕过防御方法形成攻击的检测缺陷。

1 传统解决方案概述

为了更好的解决网页篡改攻击、跨站脚本攻击（XSS）、跨站请求伪造攻击（CSRF）、电子商务网站钓鱼攻击四大Web攻击，从近些年传统的解决相关攻击的方案中寻找突破口，发现传统的关于网页篡改攻击、跨站脚本攻击、跨站请求伪造攻击、电子商务网站钓鱼攻击的解决方案存在缺陷。下面分别简单概述传统Web攻击的解决方案及其不足缺陷。

1）网页篡改检测方法。传统的网页篡改检测方法[4]有外挂轮询技术、核心内嵌技术、事件触发技术。

外挂轮询技术是利用一个网页检测程序，以轮询方式读出要监控的文件，再与真实的文件相比较。判断文件内容的完整性，对于被篡改的文件进行报警和恢复。对一般网页来说，主站及各系部的二级网站存在着内容丰富，浏览人数多且更新很快的特点。轮询扫描时间间隔大，篡改检测不及时并占用系统资源较大。

核心内嵌技术是指当外网提出修改校园网网页的请求时，必须先经过加密验证。此种方式对每个对外发布的网页都需要进行完整的调查和加密验证，对网站服务器造成了较大的负载，耗时多。

事件触发技术主要采用文件系统接口来对需要修改的文件进行验证，在被修改时进行合法性检查。这种技术是将安全保障建立在“文件不可能被隐私地篡改”这种假设上，因此也没有对文件流进行任何检查，所以在一些情形下，用户是有可能访问到被篡改的文件的。

2）针对跨站脚本攻击，微软最早提出HttpOnly，使浏览器禁止页面的JavaScript访问带有HttpOnly属性的Cookie，解决了XSS后的Cookie劫持攻击。另外，输入输出检查，通过匹配XSS的特征，比如查找用户数据中是否包含了“<script>”、“javascript”等敏感字符，在一定程度上减少XSS攻击发生的概率。然而一旦攻击发生，防御方法没法解决检测问题。

3）CSRF这种攻击方式在2000年已经被国外的安全人员提出，但在国内，直到2006年才开始被关注，而现在，互联网上的许多站点仍对此毫无防备。验证码是CSRF的一种重要防御方法，很多时候，出于用户体验考虑，网站不能给所有的操作都加上验证码。因此，CSRF攻击仍大量存在，且危害巨大。

目前全球通过最高级别的SSL证书来有效防范钓鱼攻击，通过全球可信的CA给网站颁发EVSSL证书，激活浏览器绿色地址栏，保证客户和网站之间的通信不被窃听，并醒目表明网站自己经过认证之后的身份。但就算使用强式加密的SSL服务器认证，要侦测网站是否仿冒实际上仍很困难。

2 中小型网站智能安全检测的关键技术

针对上面提出的Web攻击检测面临的诸多问题，由此设计中小型网站智能安全检测系统，系统功能如图1所示。

图1 中小型网站智能安全检测系统功能图

该系统将从网页篡改攻击、跨站脚本攻击、跨站请求伪造攻击、电子商务网站钓鱼攻击四个模块介绍该系统的关键技术和原理。

2.1 网页篡改检测

利用爬虫技术爬取待检测网站的架构和关键信息，通过网页结构分析、信息智能识别等技术不同时段对文件的校验和篡改规则匹配，判断黑客对网站的非法篡改。

原理：网页篡改检测实现的关键是对网页文件进行校验。首先对需要保护的网页文件列表进行散列函数处理，提取信息，生成用于比较的文件报文摘要校验数据。然后，根据用户自定义的安全策略，定期将这些文件的报文摘要与原始文件的报文摘要进行比较，并根据篡改规则判断是否存在篡改威胁，从而选择报警或者执行特定操作（自动恢复、不做处理记录日志文件等）。网页篡改检测模块可以利用爬虫高效爬取网页架构、读取网页信息、筛选网页特

征属性，对网页错误信息进行日志记录、信息实时比对等功能，达到解决网页信息篡改、信息增减、恶意代码攻击、插入恶意链接等攻击行为的目的。

在该检测模块内，爬虫技术的内部机制为线程池机制[5]。线程池属于对象池，最重要的特征是最大程度利用线程，并使编程模型更清楚、更优化。其原理流程图如图 2 所示。

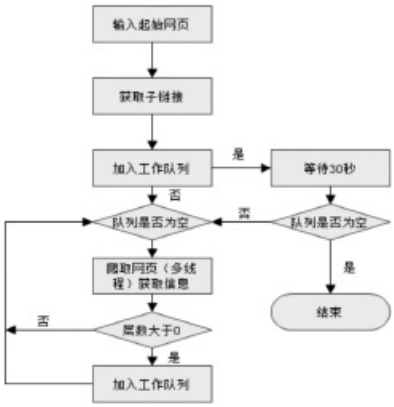

图2 爬虫技术原理流程图

在线程池机制下，递归实现对多层深度网页的访问，并提取相关网页特征属性，如字段大小、最后修改时间、正则提取架构、特征词、hash 值等。并设置 depth 深度变量，每爬取一层，深度减一，防止爬虫从子链接中又爬回原始网页形成死循环。

2.2 XSS跨站攻击检测

运用正则表达式过滤非法跨站请求，实现跨站脚本攻击检测。可以筛选过滤非法“<iframe>”跨域请求，以及对 Http 头部 Refere 来源的判断，可以扫描到被浏览器误认为合法的跨站请求。

原理：运用正则表达式过滤非法跨站请求，实现对 XSS 攻击特征字符的检测。首先，为达到对语境的理解更加深入的目的，结合出色开源的 XSS Fliter 方案，采用多种识别技术（如关键字匹配、返回信息智能识别等）并利用网络爬虫自动分析 Web 应用程序，寻找可疑的攻击语句，进行恶意攻击的判定；其次，使用网页动态参数判定和网页结构分析技术，有效地过滤非动态参数，提高检测效率；最后，再对检测到得可疑点逐一进行测试，将爬虫搜索到的可疑 URL 绑定，进行检测，减少漏报的概率。

在该模块中，允许其绑定 URL 进行检测，减少漏报的概率。其主要过程就是逐一取出测试语句进行测试，直到确认该页面存在 XSS 攻击代码或者所有测试语句取遍为止。

2.3 CSRF跨站请求伪造检测

结合 Http Referer 检测与同源策略[6]检测，确定是否有跨站请求伪造攻击。

原理：主要运用 Http Referer Check 检测跨站请求伪造攻击，实现对 CSRF 伪造的跨网站请求的检测。首先，采用多种识别技术（如关键字匹配、返回信息智能识别等）并利用网络爬虫自动分析 Web 应用程序，检测非同源的访问请求，同时结合系统中的字典寻找可疑的攻击语句，进行恶意攻击的判定；接着，对含有可疑攻击语句的页面进行 Http Referer Check，判断是否真的存在跨站请求；其次，使用网页动态参数判定和网页结构分析技术，有效地过滤非动态参数，提高检测效率；最后，再对检测到得可疑点逐一进行测试，将爬虫搜索到的可疑 URL 绑定，进行检测，减少漏报的概率。

本模块对于跨站请求伪造攻击检测也吸取了同源策略中的内容。当系统扫描整个网站发现非同源的访问请求时，就会视为恶意请求，并对其进行更进一步的检测。

2.4 钓鱼攻击检测

钓鱼攻击检测：源码层检测网站留言型钓鱼攻击。对钓鱼页面与钓取管理员账号的恶意代码进行检测，保证网站账号安全。

3 测试与结果分析

针对上述研究内容，本系统进行了系统测试与分析。本系统将测试分为四大部分：网页篡改功能测试、跨站脚本攻击检测、电子商务网站钓鱼攻击检测功能检测和跨站脚本攻击检测功能检测（在此由于篇幅原因，重点展示网页篡改功能测试，跨站脚本攻击检测）

1）网页篡改功能测试。

测试结果与分析：经过为期两月的测试，统计出 9 所高校校园网站的测试数据（如表 1 所示）。各高校网页篡改检测测试结果显示出现的网页错误类型具体集中在 UnicodeEncodeError、URLError、BadStatusLine、Timeout[7]。其中，少数校园网站的个别链接存在架构问题导致返回特殊错误。URLError 错误出现数目较多，导致此原因的 URL 错误包括多种类型，并返回网页错误代码。例如：Error 401，访问被拒绝；Error 404，无法请求 Web 页面或页面不存在；Error 504，网关超时等。管理员可以通过相应的错误类型修改和更新自己的网站，无篡改现象发生。

表1 9所高校的测试数据统计表

错误类型 / 学校名称	UnicodeDecode Error	UnicodeEncode Error	URL Error	HTTP	BadStatusLine	Timeout
北京大学	0	0	0	0	1	0
清华大学	0	2	16	0	0	2
中国农业大学	0	8	168	0	1	4
北京语言大学	0	4	14	0	0	0
中央美术学院	0	3	17	0	0	0
北京邮电大学	0	3	17	0	0	0
中央财经大学	0	1	26	0	0	2
北京科技大学	0	7	103	0	1	2
北京师范大学	0	0	20	0	0	8

2）脚本攻击检测功能测试（这里重点展示跨站脚本攻击检测）

在本次测试单元主要针对电子商务网站的 CSRF、XSS 和钓鱼等攻击，做犯罪过程重现。跨站脚本攻击检测测试结果如图 3 所示。

图3 跨站脚本攻击检测测试结果

按照攻击方式分类，XSS 可分为三类：反射型 XSS、存储型 XSS、DOM Based XSS[8]。

反射型 XSS：只是简单地把用户输入的数据“反射”给浏览器。黑客往往需要诱使用户“点击”一个恶意链接才能攻击成功。反射型 XSS 也叫做“非持久型 XSS”。

存储型 XSS：会把用户输入的数据“存储”在服务器端。这种 XSS 具有很强的稳定性。

DOM Based XSS：通过修改页面的 DOM 节点形成的 XSS 称为 DOM Based XSS。

通过我们查阅资料和对测试记录的分析，可以得出针对中小型网站，最常遭遇的是 GET 型与 POST 型的 CSRF 攻击，它们的攻击的普遍特点是都没有脱离突破 Web 层面上的一个非常重要的安全策略（即同源策略），这个策略用来限制客户端脚本的跨域请求行为，但实际上由客户端 HTML 标签等发出的跨域 GET 请求被认为是合法的，不在同源策略的限制之中，而现在的跨站攻击正是利用这个机制，比如嵌入第三方资源（图片、JS 脚本等）进行跨域请求行为。

为防止此类攻击，系统从非法请求开始前便对请求源 Refere 进行判断和身份认证，并通过对网站过滤敏感标签、后台安全机制检测和日志分析等不同方面，针对 CSRF 留言攻击进行渗透测试和防御测试，效果良好，能够有效检测出网站的安全威胁。

3）电子商务网站钓鱼攻击检测功能检测

根据钓鱼攻击原理，针对该网站进行源码级安全防护。

4）跨站脚本攻击检测功能检测

根据 CSRF 原理，针对该网站进行源码级安全防护。

3）和 4）检测方式与 XSS 类似，由于篇幅原因不在此展示。

4 结束语

本文结合网络及其安全相关协议、python 编程语言、正则表达式，从网页特征的存储与对比方面设计一个每天定时运行目标计划任务进行网页内容的抓取对比的智能安全检测系统。该系统主要面向对象为基于 Windows、Linux 系统的中、小型信息发布网站并主要针对 Web 应用服务器进行项目测试和研究，其创新点如下：

1）基于广度优先算法，设计并实现线程池机制下的快速网络爬虫。该爬虫在 10 分钟高效爬取 800~1000 条链接，并删选重复链接、父层链接避免重复循环；

2）运用正则表达式过滤非法跨站请求，实现跨站脚本攻击检测。可以筛选过滤非法“<iframe>”跨域请求，以及对 Http 头部 Refere 来源的判断，可以扫描到被浏览器误认为合法的跨站请求，从而大幅减少漏报；

3）从代码层检测网站留言型钓鱼攻击。第一时间为网站发现黑客的钓鱼留言并给予管理员提示警告；

4）在远程服务和云安全模式下，尽最大可能让用户安全得到全面保障。将获取的互联网中恶意链接、恶意关键字等的最新信息上传云端服务器，并由系统统一安排调度，供所有站点检测时使用。（责编 吴晶）

参考文献

[1] 张树壮，罗浩，方滨兴．面向网络安全的正则表达式匹配技术 [J]. 软件学报，2011，22（08）：1838-1854.
[2] 曾浩．基于 Python 的 Web 开发框架研究 [J]. 广西轻工业，2011,(08):3-10.
[3] 章明．Web 应用程序客户端脚本安全技术研究 [D]. 上海：上海交通大学，2012.
[4] 孔辉．一种网页防篡改系统的设计与实现 [D]. 北京：北京邮电大学，2011.
[5] David M.Beazley 著．谢俊，杨越，高伟译．Python 参考手册 [M]. 北京：人民邮电出版社，2011:334-364.
[6] 康海燕，陈然，苑晓姣等．基于 Android 防火墙日志系统的研究与实现 [J]. 北京信息科技大学学报，2012,27（4）：7-11.
[7] Magnus Lie Hetland 著．司维，曾军崴，谭颖华译．Python 基础教程 [M]. 北京：人民邮电出版社，2010:243.
[8] 钟晨鸣，舒少培．Web 前端黑客技术揭秘 [M]. 北京：电子工业出版社，2013:150.
[9] 苑晓姣，康海燕．一种查询日志匿名化算法 [J]. 北京信息科技大学学报，2013, 28(5): 24-27,31.