ISSN: 2661-3719(Print); 2661-3727(Online)

一种基于数字签名的高性能网站页面篡改监控系统

李　兰　张　晶
天津开放大学　天津　300000

摘　要：本文设计了一种基于数字签名技术的网站页面篡改监控系统，系统采用网络爬虫技术获取监控网页的快照并进行数字签名，将哈希值存储进非关系型数据库中，篡改检测时使用数字签名进行比对，使得系统性能得到提升。同时，系统采用模块化设计，可通过分布式部署，实现系统按监控业务量的灵活配置。此外，系统通过多管理员的赋权管理机制，可实现站点管理员和监控网站多对多的监控管理，适合于组织关系复杂的各企事业单位、高校和政府机关的网站监控工作，作为网络攻击事后发现的技术防护手段，在网络安全防护体系中发挥重要的实际作用，具有相当的应用价值。

关键词：数字签名；网络爬虫；防篡改

A high performance web page tampering monitoring system based on digital signature

Lan Li, Jing Zhang
Tianjin Open University Tianjin 300000

Abstract: This paper presents a website page tampering monitoring system based on digital signature technology. The system employs web crawlers to capture snapshots of monitored web pages and applies digital signatures to generate hash values that are stored in a NoSQL database. During tampering detection, the system compares the digital signatures for verification, thus improving overall performance. Additionally, the system is designed with modularity, allowing flexible configuration based on the monitoring workload through distributed deployment. Furthermore, the system incorporates a multi-administrator authorization management mechanism, enabling multiple administrators to manage monitoring for site administrators and monitored websites in a many-to-many relationship. This design is suitable for website monitoring tasks within organizations with complex hierarchies, such as enterprises, institutions, universities, and government agencies. As a post-detection technique for network attacks, this system plays a crucial role in the overall network security defense system and holds significant practical value.

Keywords: digital signature; Web crawler; Tamper proof

一、系统提出的背景

1. 系统建设意义

随着信息技术在各行业中的不断应用，互联网已经融入到现代生活中的方方面面，网络安全问题也被随之提升到了一个前所未有的高度。习近平总书记曾多次指出，没有网络安全就没有国家安全；过不了互联网这一关，就过不了长期执政这一关。由此可见，守好网络阵地，筑牢安全防线将是各行业在今后的信息化实践中相当重要且需长期坚持的工作。

网站页面篡改是一种非常常见的网络攻击，一旦实施成功，影响可能十分恶劣。攻击者不仅可以通过非法篡改页面发布谣言和虚假信息，而且可借此实施网络诈骗，甚至煽动舆论，严重影响网络意识形态安全。

目前，业界各种页面防篡改产品作为该攻击手段的事前防护措施，已经得到了较好的应用和推广。然而，页面篡改的事后监控措施仍然未得到足够的重视，对于一些藏匿较深的页面，往往被篡改以后相当长的一段时间才被运维人员发现，而此时，有害信息可能已经传播，恶劣影响可能已经造成。网站管理者不得不将大部分精力放在网站页面的监控上，以便在发现网站被篡改或者访问故障时第一时间处理，浪费了大量的人力资源。然而，单纯靠人工力量只能监控关键页面，而整个站点页

课题项目：本文为天津市教育科学规划课题－“终身教育环境下网络安全体系研究”（课题批准号：HCE210333）研究成果

作者简介：李兰（1982年5月），女，辽宁省辽阳市，汉，中级工程师，本科，网络安全和信息化。

面数量庞大，靠人工监控几乎不可能，且通过人工监控无法保证页面变化快速被发现。因此，设计并应用一个能对大量网站页面进行监控并在页面发生变动后迅速通知运维人员的系统就显得尤为重要。

2.相关研究

近年来，网站页面篡改监控技术领域的研究非常广泛，很多学者进行了较为深入的研究。内蒙古大学计算机学院的阮宏玮等人提出了一个基于快照轮询和文本检测的批量网页防篡改系统[1]，该系统通过定义网页中的比对部分，以网页快照的形式进行文本比对检测，满足了网关人员的网页监控需求。上海市信息安全测评认证中心的陆昊则基于云架构，设计并实现了一个基于云架构的网站防篡改在线检测平台[2]，该平台采用Hadoop分布式数据处理架构，借助Hadoop文件分布式系统和映射违约算法，实现了具备弹性和高可用性。在恶意篡改方面，该系统分析篡改后的文本特征并结合网页变化时间等因素，并配置权重值，通过聚类算法计算阈值，在页面是否为恶意篡改的分析上进行了尝试。在高校网站群监控领域，云南师范大学王宁邦等人，利用网页爬虫技术，构建了高校门户网站群网页的篡改预警监控系统[3]。该系统使用主题爬虫技术，并对爬虫任务进行切分，可以通过邮件、短信和微信等多维联动的告警机制告知站点管理员。此外，针对高校门户网站安全方面，王宁邦等人以信息安全等级保护规范作为指导，结合该预警检测系统，形成了校园网络安全的联动工作机制和方法，具有一定借鉴意义和参考价值。

在网站恶意篡改检测方面，武汉邮电科学研究院的李柯言等人提出了一个基于特征识别的网页篡改检测系统[4]，避免了HASH校验和DOM树结构对比的缺点，使用网络爬虫对防护页进行截图和实时流量爬取并获取爬取时间，设定检测强度，再由敏感词识别器，违规图片识别器进行识别，最后将识别器数据输入至分析模型中对是否发生网页篡改进行判断，并输出最后的分析结论，解决了原有检测方法误报率较高等问题。

本文介绍的网站页面篡改监控系统定位于各企事业单位、高校和政府机关等内容发布信息系统，一方面，通过可配置的正则表达式，过滤页面中的动态内容；另一方面可进行分布式部署，结合非关系型数据库，提供高性能、高可用性的页面篡改监控服务。此外，系统使用可调度的爬虫算法，进行广度优先的页面爬取，优化了页面爬取速度，增强了系统的承载能力。

3.系统设计目标

本文介绍的网站页面篡改监控系统为网络管理工作中费时费力的网站页面篡改监控工作而设计。管理员配置好各项监控参数后，由专用的服务器代替人力对监控网站的页面进行访问，爬取页面内容，访问间隔可人工设置。一旦发现所监控的页面出现变动则会以邮件、短信或APP推送等形式发出警报，由运维人员确认页面变动是否为正常修改，正常修改可确认警报，被非法篡改则及时处理。

网站页面篡改监控系统，在整个网络安全防护体系中作为页面防护的最后一道屏障，应能满足以下要求：

首先是高可靠性方面的要求，页面篡改威胁没有特殊的时间特性，系统应具备7X24小时的不间断运行能力。

其次是高性能方面的要求，对于一般单位的门户网站，系统应能监控万级页面状态，对于拥有二级单位或使用站群系统的大型单位，系统应能承载十万级甚至百万级页面的监控任务。

再次是响应时间方面的要求，在页面发送变化后，系统应能在秒级或分钟级检测到页面变化并迅速通知运维人员。

最后是对于监控网站页面适应性方面的要求，应能适应含有部分动态内容（如当前日期、点击数、浏览数或点赞数等）的页面监控，而不会因为其正常变化的动态部分频繁触发报警。

二、系统功能设计

1.系统角色划分

网站页面篡改监控系统按其各个业务模块的定位，将系统功能划分为两个角色，分别为系统管理用户和站点运维用户。

2.系统管理用户功能设计

系统管理用户负责整个系统账户的维护、监控网站的管理、运维用户与监控网站的权限分配以及系统设置。

（1）账号管理

系统管理用户可通过账号管理功能创建站点运维用户，并可进行站点运维账号的开启、关闭、重置密码和修改信息等账号管理操作。

（2）监控网站管理

通过监控网站管理功能，管理员可配置监控站点的各项信息。配置信息包括站点URL、分配爬虫最多数量、爬虫扫描频率、爬虫爬行深度、监控时间段、监控保障优先级、监控日志保留级别、是否监控页面异常状态等。

通过设置“是否监控页面异常状态”，网站页面篡改监控系统可以在监控页面篡改情况的同时监控页面异常状态，包括页面无法访问或异常的HTTP状态码（HTTP

状态码不为200的情况），使运维人员第一时间发现网站故障或者死链（404 not found）。

在配置监控站点信息时，如果监控网站内含有动态数据（如系统日期、点击次数、验证码等），管理员需要正确配置正则表达式告知系统页面内的动态部分的数据特征，网页爬虫在每次获取到页面时，会按照正则表达式设置的规则去除页面动态部分，防止由于动态数据频繁触发误报警。

（3）权限管理

系统管理员通过权限管理模块为站点运维账号分配监控站点，运维账号和监控站点之间的关系为一对多关系，即一个运维账号可接收并确认多个站点的报警信息以及相关监控日志的查看和管理。

在进行权限分配时，可指定某个网站的报警信息以何种方式通知站点运维用户，报警通知手段包括监控大屏显示、电子邮件、手机短信和APP推送。

（4）系统设置

系统设置模块为整个系统提供了参数配置功能，包括发送系统邮箱账号配置信息、短信网关配置信息、数字签名算法等。

3.站点运维用户功能设计

（1）个人信息管理

站点运维用户登录系统后，可通过个人信息管理修改个人的账号信息，如登录密码，手机号码，电子邮箱等，也可设置自己首选的告警接收方式，系统将优先以运维用户自己设定的告警方式推送告警信息。

（2）查看/确认告警信息

站点运维用户所负责的监控站点在监控时间段内发生变化后会触发告警事件后，站点运维用户可通过该功能查看告警信息。告警信息中会附带发生变化页面的URL，用户可直接跳转到相关页面以查看该页面是否为正常更新或被非法篡改。处理完毕后可通过该模块确认报警信息。

（3）告警日志管理

站点运维用户可通过告警日志管理功能模块浏览、查询过往的告警日志，实现篡改记录可追溯。发生页面篡改事件时，该记录可作为网络攻击行为证据链中最后一个日志证据。

（4）站点监控统计分析

站点运维用户可通过站点监控统计分析功能，对监控网站的页面进行分类统计分析，分析爬取的页面总数、页面平均长度、页面平均爬取时间、页面的变化频次、异常频次等信息。

三、系统架构设计

1.系统架构

本文介绍的高性能页面篡改监控系统分为UI模块、线程调度器、页面爬虫、页面分析器、篡改报警器、日志模块和底层的Redis与MySQL数据库。整个系统的架构如图1所示。

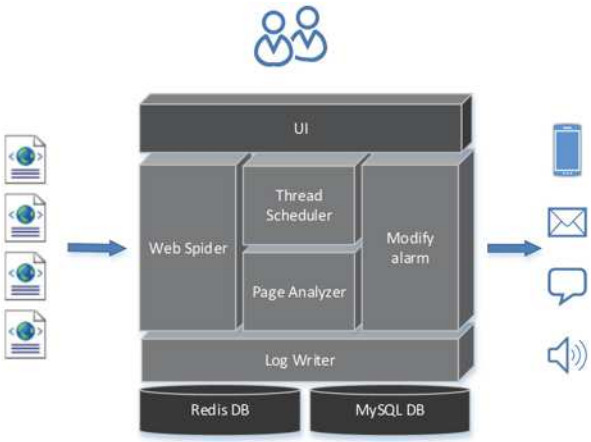

图1　系统架构图

UI模块采用B/S架构，客户端部分使用HTML5+CSS3，负责与管理人员进行交互，服务器端将配置信息存储于关系型数据库中（MySQL）。管理员可通过UI模块对系统进行网站配置、监控频率配置、报警设置、查看日志和账号权限管理等操作。

线程调度器负责整个系统的线程管理，如开启及回收爬虫线程，启动页面分析线程，启动报警线程等。线程调度器同时监控系统CPU和内存资源，当系统资源出现不足时，线程调度器可能会强行回收深度较高的爬虫，释放系统资源，保证高优先级任务（如启动报警线程）得到执行。

页面爬虫根据配置的站点信息，按照广度优先的算法进行页面爬取，爬取深度由管理员配置控制，可为1到10级。一般中小型网站设置一个爬虫进行爬取，对于大型网站，可通过修改配置开启多个爬虫，按广度分担爬取任务。

页面分析器负责页面静态内容的提取并进行哈希，计算其数字摘要，查找非关系型数据库（Redis）中已存的摘要并进行比对，发现摘要不同后，通知篡改报警器进行报警处理。

篡改报警器按照管理员设置的通知方式，以邮件、短信、APP推送或Web界面报警等方式通知运维人员，告知其发生变化的页面地址、发现时间，由运维人员确认报警信息。

日志模块负责记录整个系统的运行日志、管理员

操作日志、报警日志，并通过UI界面提供日志的查询和管理功能。日志以记录形式存储于关系型数据库中（MySQL）。

Redis数据库和MySQL数据库为系统提供底层的数据库支持。Redis作为高性能的key-value数据库，非常适合存储页面摘要信息。而MySQL作为关系型数据库，则适合存储用户权限、系统配置和日志记录等数据。

2.系统工作原理

管理员登录系统UI界面后，首先对要监控的网站进行配置，设置爬虫个数与爬行深度、监控频率、监控时间段、动态内容正则表达式、通知告警方式、监控优先级等信息。启动监控后，线程调度器会启动页面爬虫，按照用户配置开始以广度优先算法爬取页面。

页面分析器得到爬取的页面以后，通过配置的正则表达式，过滤掉页面中正常变化动态内容（HTML标签），如当前日期、页面点击次数等，对其余静态内容和页面URL地址分别进行哈希，计算数字摘要。然后，页面分析器以URL地址的数字摘要为主键，在Redis数据库中查找已存记录。如果未找到记录，则表示该页面第一次被爬取，页面分析器会将该页面的URL地址的数字摘要和静态内容的数字摘要以key-value的形式存储于Redis数据库中，同时记录爬取时间。如果找到记录，页面分析器则比对静态内容的数字摘要和已存记录是否一致，若二者一致，则表示页面未发生变化，仅更新爬取时间，若二者不一致，则说明该页面被篡改。

页面分析器在监控时间段内发现页面被篡改后会通知篡改报警器。篡改报警器向线程调度器申请启动报警线程，并根据用户设置的报警方式，执行报警任务，通知运维人员。

运维人员在接收到报警信息以后，可直接点击报警信息中的链接，查看被篡改的页面是否为恶意篡改或正常更新。若发现为恶意篡改，则可迅速处理；若为正常更新，则可直接确认报警信息。警报确认后将，页面分析器会将变化后的摘要存储于Redis数据库中，作为下一次比对的依据。

页面爬虫在爬取完整个网站后，会将爬取到的所有页面地址、爬取时间和版本信息存储在Redis数据库中。在一段时间内，爬虫不会再次重新爬取整个网站，而是根据这些信息直接提取页面信息并进行比对。同时，若发现URL地址返回HTTP状态码为404（未找到）时，爬虫也会在Redis数据库中删除对应页面；对于URL地址返回HTTP状态码500（服务器内部错误）类型或网站无法连接时，爬虫也可根据配置信息直接通知篡改报警器，提示运维人员服务器故障。

四、结语

本文设计的基于数字签名的高性能网站页面篡改监控系统，已经由天津开放大学网络安全与信息化办公室进行开发完成。系统采用HTML 5、CSS3和.Net framework 4.0技术实现，页面爬虫基于python 3.7.1开发，系统部署于H3C虚拟化集群中。系统在2021年、2022年的重要时间节点的网络安全保障工作以及各级各类网络攻防演练中发挥了重要作用。

相对于同领域的其它研究相比，本文设计的网站页面篡改监控系统更加注重于系统承载性能，系统放弃使用全文比对，使用数字签名并结合广度优先的爬虫算法，尽可能多地保护站点页面，在页面发生篡改后，尽可能迅速地进行报警响应。在业务上，系统也支持多管理员的权限划分和监控任务分工，更适合组织关系复杂的中大型企业、高校或机关单位的网站监控需求。

然而，目前系统的主要缺点也在于为了提升性能而使用数字签名进行比对，无法对内容进行深入的大数据分析和人工智能识别，因此虽然系统的恶意篡改报告率为100%，但存在较高的误报率，适合于更新不频繁的信息发布类站点。针对此问题，今后可将系统的页面分析器模块进行重新设计，采用基于内容提取与分析的设计思路，优化内容分析算法，结合相关领域的大数据分析或人工智能相关研究成果，降低系统误报率。

参考文献：

[1]阮宏玮，李华，王小雨，吴承勇，庞滨.基于快照轮询和文本检测的批量网页防篡改系统[J].广西大学学报（自然科学版），2011，36（S1）：142-147.

[2]陆昊.基于云架构的网站防篡改在线检测平台设计与实现[J].软件产业与工程，2015（05）：27-30.

[3]王宁邦，徐博.基于爬虫和网页防篡改的高校门户网站群预警监控系统构建[J].云南民族大学学报（自然科学版），2019，28（05）：502-509.

[4]李柯言，刘晓东.基于特征识别的网页篡改检测系统[J].电子设计工程，2020，28（18）：16-19+24.