

一种高性能 Web 内容安全监管系统

张玲¹ 王喜¹ 张立¹
ZHANG Ling WANG Xi ZHANG Li

摘要

Web 内容安全保障主要是保护大型政府门户网站、新闻网站发布的公告和新闻等内容的安全。这类 Web 网站发布的内容是社会舆情引导的重要组成部分,如果被恶意篡改将会引发严重的后果。针对这一类问题,提出一种 Web 内容安全监管系统。系统从内容监管、权限审核、核心词表与关联模式定义三方面实现对 Web 内容审计与管控,融合了被动与主动两种安全保护的方式,并能与内容发动审核流程紧密联合。所提出的系统相对于其他同类系统具有高性能、高安全保障性的特点。

关键词

Web 内容安全; 内容监管; 高性能; 监管系统; 门户网站

doi: 10.3969/j.issn.1672-9528.2024.01.035

0 引言

目前我国社会信息化程度已相对较高,绝大多数的政府部门、事业单位都有自己的门户网站。这些网站会发布一些非常重要的公众信息,包括各类公告、宣布事项等。这些网站是社会公众及时了解事态、权威信息的重要渠道,也是社

会宣传、引导社会舆情的重要组成部分。这类网站发布的内容具有重要的社会意义,必须要保证其发布内容的权威性与可靠性。但是,经常看到这些网站受到黑客的攻击,或由于发布审核流程出现问题,引发舆情事故。分析这类事故产生的原因,主要有以下三个方面。

(1) 内容未及时更新与维护。大部分门户网站经过长期的运营,积累了大量的信息内容,如果网站的内容没有建

1. 泰州市气象局 江苏泰州 225300

5 结语

无线传感器网络的信息隐藏算法是保护数据安全的重要手段之一,本文在边缘计算场景下研究了无线传感器网络信息轻量级加密算法。首先通过贝叶斯网络对网络信息进行分类,从而识别出敏感信息作为待隐藏信息;然后再利用轻量级加密技术对待隐藏信息进行了加密处理,并将加密后的信息嵌入到常规载体信息中;最后,设计了敏感信息随机嵌入的智能隐藏算法,使得敏感信息在正常情况下无法被察觉,有效防止了未经授权的访问和窃取。

参考文献:

- [1] 罗佳,周日贵,李尧翀,等.基于量子生成对抗网络的信息隐藏方案[J].计算机辅助设计与图形学学报,2021,33(7):983-990.
- [2] 秦川,李蓉受,钱振兴,等.基于宋词生成的大容量构造式信息隐藏算法[J].计算机学报,2023,46(1):17-30.
- [3] 张敏情,王泽曦,柯彦,等.基于多项式秘密共享的图像密文域可逆信息隐藏[J].电子与信息学报,2022,44(12):4337-4347.
- [4] 龚健虎,张跃进.深度 AWB 结合改进 DIT 的高效大数据分类[J].计算机工程与设计,2021,42(2):468-474.

- [5] 吴家皋,郭亚航,蔡沈磊,等.基于多时间段优化贝叶斯网络的车载容迟网络路由算法[J].通信学报,2021,42(12):109-120.
- [6] 刘政,代培培,邢建平,等.面向嵌入式设备的轻量级 AES 加密通讯系统的设计与实现[J].电子器件,2023,46(1):29-35.
- [7] 李玮,刘春,谷大武,等.Saturnin-Short 轻量级认证加密算法的统计无效故障分析[J].通信学报,2023,44(4):167-175.
- [8] 温文嫒,邹孟雷,方玉明,等.面向医学图像加密域大容量信息隐藏与认证方法[J].中国图象图形学报,2023,28(3):716-733.
- [9] 吕海翠,郭玲.数据库敏感信息可变保序加密安全算法研究[J].计算机仿真,2021,38(9):301-305.
- [10] 罗宇,郭家松.大位宽情况下的回滚式循环冗余校验算法[J].电子与信息学报,2021,43(4):1057-1063.

【作者简介】

邓喆(1989—),女,贵州铜仁人,硕士,讲师,研究方向:计算机应用技术、大数据技术、网络安全。

吴亚洁(1983—),女,天津蓟县人,硕士研究生,讲师,研究方向:电子信息。

(收稿日期:2023-10-17)

立相关索引，则在需要变更时，易于出错和漏失，如：人员发生变动，但是早期的相关内容还没有被撤换。

(2) 缺少内容监管。虽然大部分政府、组织对于重要公告内容的发布都制定了相关的审核流程，但是这些流程并没有与 Web 网站建立数据操作上的关联，因此无法直接对 Web 内容发布实施管控。工作人员无意中篡改网页或者修改了网页中链接指向资源的位置，都会造成信息的非法发布或资源失效，如：日本埼玉县政府官网操作失误，泄露 191 名新冠患者信息事故等^[1]。

(3) 缺乏动态监控。网站运营过程中，由于系统自身的原因可能会造成链接失效、无法访问，或者由于黑客攻击，如挂马、恶意篡改内容等造成严重的事故，而现有的 Web 系统缺乏对这类事件的监控^[2]。

为了应对上述问题，Web 内容安全监管系统应运而生。与传统的 Web 安全保障系统侧重点不同，传统的 Web 安全保障系统注重于系统、软件与服务的安全性、可靠性，而 Web 内容安全主要注重于 Web 发布内容的完整性、可信性。目前 Web 内容监管主要分为被动监控技术和主动监控技术。被动监控技术是 Web 服务器上安装监控程序，在用户访问内容的过程中，进行信息审核、过滤^[3]。而主动监控技术^[4]是指监控系统周期性地利用爬虫对目标门户网站上发布的内容，进行抓取、分析、审核，以期发现黑客攻击痕迹、无效链接等。显然，无论是主动与被动监控都不能完全解决前述的 Web 内容安全问题，是因为这些系统没有与内容发布审核流程相关联，也存在效率方面的问题。

为此，本文提出一种高性能的 Web 内容监管系统，与同类系统相比，不光能实现内容监控，更与 Web 内容发布审核的流程相结合，实现对内容发布的管控，具有内容索引、查询与高性能的特点。

1 系统架构与实现

1.1 系统整体架构

本文所提的 Web 内容监管系统整体架构如图 1 所示。



图 1 本文系统的部署图

从图 1 可以看出，本文所提系统的部署位置位于 Web 内容服务器与用户之间，实现了用户与 Web 系统之间的安全隔离。外部用户在访问 Web 服务提供的内容之前会经过本文系统转发请求，将请求映射到内容服务器对应资源位置。Web 服务器回复请求的内容再经过本文系统过滤之后，转发送至

外部用户。从本文系统的部署可以看出，该系统是作为一个独立的监管系统运行于 Web 服务器之外，这样使得本系统可以兼容现有 Web 内容服务系统，不会对已有系统的运行产生干扰。此外，也能隐藏本监管系统，避免暴露而受到黑客的攻击。

本文系统的架构总体上可以分五层。

(1) 配置层。配置层是整个系统的最底层，主要功能是实现系统的全局参数配置与管理。为了尽可能地减少本文系统对所保护的系统正常运行的干扰，减少对其服务性能的影响，所以本文的配置层主要配置服务请求路由的转发规则，实现对所保护系统的资源隐藏。

(2) 路由层。这一层主要是执行底层配置层中已经配置好的路由转发规则，是内容资源保护的执行层。

(3) 数据层。该层是整个系统的数据存储与管理层，其中主要包括的数据有用户数据，这里的用户系统是指维护与审核 Web 内容的用户，包括用户信息与权限。关键词索引数据，即 Web 内容关键词的索引数据，用于内容需要变更或调整时，对全 Web 内容进行检索，以防漏失。内容哈希数据，是实现对 Web 内容的全局性哈希存储，是防止发布内容被篡改，审计内容变化的关键数据。

(4) 业务层。该层是基于配置层和数据层实现具体的业务功能，如图 2 所示，主要有用户权限管理模块、Web 内容访问日志、内容 VCS 版本管理、检索与批量更新以及内容审核业务流等。其中，Web 内容访问日志模块主要用于统计外部用户的访问情况，计算 Web 服务的压力。该层的具体功能可以根据业务需求进行扩展。

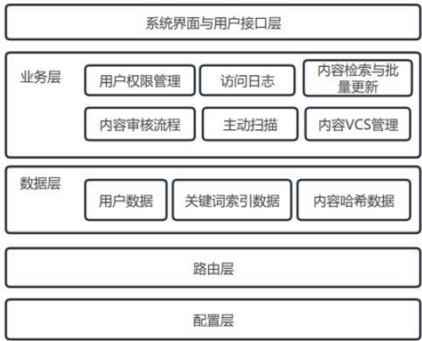


图 2 高性能 Web 内容监管系统整体架构图

(5) 系统界面与用户接口。这一层是外部接口与用户操作的界面，是整个系统的用户层。

1.2 系统核心实现技术

1.2.1 基于 Merkle 树实现的 Web 内容 VCS 管理

Web 内容版本控制系统 (version control system, VCS) 管理对于 Web 内容安全具有重要的作用，一方面通过 VCS 管理可以发现 Web 内容的修改过程，追踪溯源 Web 内容的

发布与操作日志,为 Web 内容审计奠定基础;另一方面 VCS 也可以及时地发现 Web 服务受到了黑客攻击被篡改的部分,如 Web 服务被插入恶意代码、Web 网页被挂马等能引发严重后果的安全事故。

作为监管系统实现内容 VCS 管理与普通的 VCS 管理系统^[5]有一定的区别,主要体现在效率上。因为监管系统一方面要周期性的扫描 Web 发布的内容,另一方面需要在用户访问时拦截访问的内容并做 VCS 校验,因此内容监管系统的 VCS 校验在效率上的要求相对于普通的 VCS 系统要高很多。为此,本文提出使用 Merkle 树^[6]实现对 Web 内容的 VCS 管理。

如图 3 所示, Merkle 树是一种二叉或多叉树的数据结构。Merkle 树具有一些优良的特性,如快速校验、定位校验出错位置等,在区块链等众多应用中都有使用。

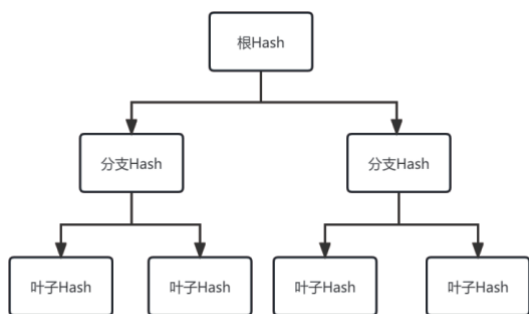


图 3 Merkle 树结构

本文使用 Merkle 树结构来对 Web 内容实现 VCS 版本管理,其实现的基本原理是:首先对用户发布的 Web 内容生成要校验的数据分块,即将要发布的 Web 内容文本切分成段,然后对这些分段进行 Hash 运算,最后使用分段的 Hash 值分层构建生成 Merkle 树,并存储至上文所述的内容哈希数据库中。具体的过程如表 1 所示。

表 1 内容校验 Hash 生成过程

1. 通过内容扫描模块,获得 Web 发布内容文本。
2. 根据本文系统配置层配置的参数,将文本分段,本文系统对于同一内容的文本分段,分别从头至尾,以及由尾至头两个方向进行分段,记为 H 和 T 。其中: H 为从头至尾的文本分段数组,而 T 为从尾至头的文本分段数组。
3. 分别针对 H 和 T 数组中的文本分段,两两分组,计算 Hash 值,形成 $Hash_H$ 、 $Hash_T$ 两个 Merkle 树的叶子 Hash 值,如图 3 所示。
4. 使用 Merkle 算法,由叶子节点的 Hash 值自底向上构建出 Merkle 树形成 $Hash_H$ 、 $Hash_T$ 两个对应的 Merkle 树。
5. 最终,本文系统会将上述的 $Hash_H$ 、 $Hash_T$ 对应的 Merkle 树、 H 、 T 文本分段数组,及其对应的内容 URL,校验生成的时间戳组成记录。并将记录插入内容 Hash 数据库中,完成内容的 Hash 索引创建。

在构建完成内容 Hash 库之后,本文所提出的系统会采用主动方式周期性扫描校验 Web 内容,在用户访问 Web 内容的时候拦截并校验 Web 内容。校验过程与生成内容 Merkle 树过程相反,同样使用并行的方式生成 Merkle 树,并逐节点进行校验,一旦发现 Hash 值与当前内容 Hash 库中存储的 Hash 值不相符,即发现内容被篡改。

由上述的流程可以看出,本文系统针对每一个 Web 发布的内容都创建两棵生成方向不同的 Merkle 树。因此,本文系统校验过程中,通过两个方向并行校验,不同方向上发现的第一个 Hash 值不同的叶子节点,即为被篡改的内容位置,从而可以准确地定位被篡改文本的具体位置。

Merkle 树结构可以按 Web 内容分段,并行执行 Hash 校验,实现高效的文本内容版本管理,因此本文采用 Merkle 树结构来实现对内容的 VCS 管理与校验。除了高效之外,还因为 Web 内容中经常会存在下载资源、图片、CSS/JS 文件等外部引用的文本或非文本类型资源,使用 Merkle 树结构可以对这类资源的 Hash 值进行兼容,因为本身 Web 内容 HTML 结构本身也是树形的结构。

1.2.2 内容审核与发布监控业务流程

内容审核是内容发布之前必要的审核批准业务流程,以保证待发布内容的正确性。一般政府部门、大型组织的门户网站的内容在发布前需要经过不同权限的人员签发同意后,才能正式发布。但是现有的大部分审核流程是游离于 Web 内容服务系统之外,采用手工办理的方式。这种方式采用手工纸质文件传递方式,一方面办理速度比较慢、效率低,另一方面对后期的审核操作追溯也不方便,此外更重要的是,这种处理流程并不能直接对 Web 内容服务进行控制,所以经常有误发布、误修改之类的事情经常发生。

为此,基于上文所述的 Web 内容监管系统框架,本文实现内容审核与 Web 内容服务的整合与关联。其实现的途径如图 4 所示。

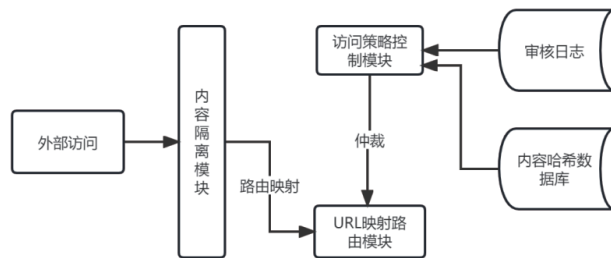


图 4 内容审核与访问监控框架图

从图 4 中可以看出,本文系统中实现内容审核与访问监控是通过路由映射和访问策略控制模块来完成的。

为了能实现对内容的实时保护,防止黑客攻击与篡改,本文系统设置了内容隔离模块。由于本文系统是部署在用户与 Web 内容服务的访问入口处,通过隔离模块拦截用户的访问。本系统对用户访问的隔离除了能实现对用户访问转发,

从而隐藏 Web 内容真正的服务器位置目标之外，更重要的是作为访问策略控制模块的执行模块来拦截未被审核通过的 Web 内容。

执行对 Web 内容审核业务流的是访问策略控制模块，该模块是通过底层的配置层来获得 Web 内容审核配置项，结合审核业务日志表来实现对外部用户的 URL 访问审核。其中，审核业务日志表是用户对 Web 内容审核的记录，该表也是用户审核业务流的基础表，表字段定义为：

<URL, ID, VCSID, SIGNS> (1)

式中：URL 是 Web 内容的访问地址，ID 是审核日志的唯一标识，VCSID 是内容版本的 ID 标识，该标识是上述 VCS 模块对于内容生成的 VCS 记录的唯一标识。显然，同一内容不同版本的 VCSID 号不一致，因此对于同一内容不同版本对应的审核日志记录也不相同。

可以通过不同 URL 模式字符串来配置对不同 URL 指向的 Web 内容并执行相应的访问控制方式。本系统 Web 内容的访问控制方式配置采用三元组的形式来表述：

<URL_Pattern, Cond, Require_Roles> (2)

式中：URL_Pattern 是指内容的 URL 模式，可以通过正则表达式来指定一批 URL，Cond 是审核通过的条件，条件使用表达式 OR、AND 来表示。Require_Roles 是需要审核的角色集合。

结合内容审核日志与访问策略控制配置，图 4 中模块执行审核的过程如表 2 所示。

表 2 内容审核过程

1. 隔离模块拦截外部用户访问之后，分析外部用户访问的目标 URL 地址，并提交到访问策略控制模块。
2. 该模块从现有的访问控制方式配置中按模式查找与访问目标区块的访问控制配置，如果没有找到对应的访问控制配置，则返回审核失败，隔离模块会拒绝此次外部用户的访问。
3. 如果有相应的访问控制配置，则取出配置，根据 Cond 配置项，构建条件表达式。
4. 再根据目标 URL，查询审核日志表，若查找不到相应的审核日志记录，则同样拒绝外部用户的这次访问。
5. 如果查找到，则取出日志中记录中的 SIGNS，构建出审核签名集合，该集合中包括同意发布签名的用户角色。
6. 针对所获取的内容审核用户角色集合，按访问配置要求，执行 OR、AND，判断对应的内容是否已经通过审核，可以正常访问。如果否，则拒绝外部用户的这次访问，如果是，则表明此次外部用户访问的内容已经经过内部审核通过，可以访问，返回审核成功至隔离模块。
7. 隔离模块在得到审核模块审核成功的返回消息后，通过底层的路由映射配置将访问地址映射到内部的 Web 内容地址，并返回给用户。

2 总结

本文提出了一种高性能的 Web 内容监管系统，该系统利用 Merkle 树建立 Web 内容的 Hash 值，从而实现对内容的版本管理。同时本系统提供了灵活的访问控制策略与执行模块，利用内容隔离模块实现对外部用户访问审核，同时校验内容是否被篡改以及是否经过审核同意后发布。

相对于同类的内容监控系统^[7-11]而言，本文系统的优点主要体现在两个方面。

(1) 使用了 Merkle 树建立了双向的内容 Hash 校验值，可以实现并行化高速的内容 Hash 校验，准确定位到内容被篡改的位置。

(2) 本文系统还整合了内容审核流程，能实现内容审核业务与外部实时访问审核联动，进一步保障了内容的安全。

参考文献

[1] 人民网. 日本埼玉县政府官网操作失误 泄漏 191 名新冠患者信息 [EB/OL].(2021-01-26)[2023-06-25].<https://m.gmw.cn/baijia/2021-01/26/1302072137.html>.

[2] 姜砚平, 杨晓东. 政府网站面临的信息安全威胁 [J]. 保密工作, 2011(4):54-56.

[3] 曲金东. 基于被动测试的 Web 性能与故障检测系统研究与实现 [D]. 广州: 华南理工大学, 2012.

[4] 耿乐群. 基于主动搜索的论坛内容监管技术研究 [D]. 哈尔滨: 哈尔滨工程大学, 2011.

[5] 龚炯, 饶若楠. DCG-VCS: 一种基于 DCG 的版本管理系统 [J]. 计算机应用与软件, 2007(8):90-91+105.

[6] 郑李伟, 王雪平. 基于 Merkle 树的安全云存储系统 [J]. 计算机系统应用, 2022,31(4):81-90.

[7] 陈红松, 胡长军. 网络内容监控与预警系统的设计与实现 [J]. 计算机工程与设计, 2009,30(1):213-215.

[8] 王永. 基于内容过滤的网络监控技术研究 [J]. 科技创新导报, 2018,15(14):147-148.

[9] 魏来. 内容汇聚子系统中监控与异常分析模块的设计与实现 [D]. 北京: 北京邮电大学, 2018.

[10] 孙楠. 网络内容与舆情监控系统的设计与实现 [D]. 南京: 南京邮电大学, 2016.

[11] 褚涛旒, 孙扬. 网络舆情监控系统处理流程及核心功能实现 [J]. 信息技术, 2022(6):12-16.

【作者简介】

张玲(1988—)，女，江苏盐城人，本科，工程师，研究方向：网络信息安全。

(收稿日期：2023-10-08)