许昌学院本科毕业论文（设计）开题报告

学生姓名	张释文	学号	5005160065
所在学院	信息工程学院	专业	网络工程
指导教师	平源	职称	副教授
论文题目	基于 OpenResty 的 Web 异常访问行为检测组件设计		

一、选题的依据和意义

Web 是一种基于超文本和 HTTP 的、全球性的、动态交互的、跨平台的分布式图形信息系统。随着 Web 应用的发展，Web 服务器逐渐成为主要的被攻击目标[1]。

过去企业通常会采用防火墙作为安全保障的第一道防线，而传统防火墙在阻止利用应用程序漏洞进行的攻击方面，却没有办法。此背景下，WAF（Web Application Firewall）应运而生。通过执行一系列针对 HTTP，HTTPS 的安全策略，来专门对 Web 应用，提供保护的一款产品[2]。WAF 初期是基于规则防护的防护设备；基于规则的防护，可以提供各种 Web 应用的安全规则，WAF 生产商去维护这个规则库，并实时为其更新，用户按照这些规则，可以对应用进行全方面的保护[3]。

但随着攻防双方的不断过招，攻击者也摸清了这一套传统的防御体系并打破了这套防线。同时这套防护思路，还有一个天生的缺陷，就是难以拦截未知的攻击。因此技术的革新也是必然的。本论文选题就是为了扩充 WAF 的功能的同时，提供高性能的解决方案。因此选用 OpenResty 平台开发一款 Web 异常访问行为检测组件。OpenResty 作为新兴平台，其是由 Nginx 核心加很多第三方模块组成，其最大的亮点是默认集成了 Lua 开发环境，使得 Nginx 可以作为一个 Web Server 使用。借助于 Nginx 的事件驱动模型和非阻塞 I/O，可以实现高性能的 Web 应用程序[13]。而且 OpenResty 提供了大量组件如 Mysql、Redis、Memcached 等等，使在 Nginx 上开发 Web 应用更方便更简单[4]。目前在京东如实时价格、秒杀、动态服务、单品页、列表页等都在使用 Nginx+Lua 架构，其他公司如淘宝、去哪儿网等也在使用[5]。

二、研究现状

目前的开源 WAF 基本采用传统服务器架构，采用性能较低的解决方案对请求数据进行过滤，技术模型并不具备非阻塞的访问网络 I/O，工作进程只能等待网络 I/O 的返回，对服务器性能有很大的影响。此外，在功能方面也仅仅是对攻击行为进行检测防御，并不能对用户的其他异常访问行为，包括恶意流量、爬虫、欺骗等操作进行分类检测。

而国内的主流 Web 应用防火墙产品，如阿里云云盾 WAF[6]和华为云 WAF[7]很好的改善了 WAF 的功能。对于反爬虫功能，云盾 WAF 提供了“精准访问控制”功能，通过检验请求 URL 和 HTTP 请求头中的 User-Agent 字段来判断是否为爬虫，并设置检测为爬虫后的应对措施[8]。华为云 WAF 提供了类似的“精准访问防护”功能[9]。此外阿里云提供了独立的“爬虫风险管理”产品，通过阿里云的爬虫数据集判断是否为爬虫，不允许自定义规则[10]。华为云 WAF 则以“Robot 检测”提供了反爬虫功能，未见能自定义规则的配置[8]。此外，云盾 WAF 和华为云 WAF 对于高频爬虫的防御措施还有通过配置 CC 攻击防护，即通过设置 IP 或 Cookie 和给定 URL (或 URL 前缀)，对超过频率阈值的用户进行封禁、限频或要求人机验证[11]。

综上所述，根据对云盾 WAF 和华为云 WAF 的研究，得出针对恶意访问 Web 应用防火墙的主要解决方案有:

(1)基于精准访问防护，设置 HTTP 请求的过滤规则，拦截恶意访问[12]。

(2)基于 CC 攻击防护，在给定 URL 下统计访问频次，拦截高频恶意访问。

(3)利用大数据和机器学习的原理，以访问特征和用户为单位进行拦截[13]。

解决方案(1)(2)的立脚点是单个 HTTP 请求或单个用户。而解决方案(3)实际上过滤的也是单个用户[14]。若需要从短时间内是否有产生大量异常行为的角度进行风险控制，从国内外的解决方案现状来看只能在 Web 应用内部完成统计，并不能单纯的借助 Web 应用防火墙来实现[15]。

三、解决的主要问题

(1) 高性能并实现非阻塞的访问网络 I/O。利用 Nginx 的特性，同时兼顾性能与功能。

(2) 针对协议请求的异常检测。对协议请求进行异常检测，检测不符合协议规格的内容。

(3) 对异常行为的信息搜集与分类。初始的规则库对异常行为进行分类，利用模型对分类后的数据集进行训练，增加准确性。

(4) 利用规则库与建立应用数据模型协同判断异常行为。

(5) 检测短时间内是否产生大量异常行为。

四、研究方法

(1) 学习 OpenResty 与 Nginx 开发技术；

(2) 设计规则库的初始状态与用户自定义接口；

(3) 学习机器学习模型算法并建立数据模型；

(4) 进行模块化编程实现；

(5) 打包整合源代码；

(6) 撰写毕业（设计）论文。

五、选题的特色及创新点

特色：

(1) 异常检测协议。对协议请求进行异常检测，检测并拒绝不符合标准的请求。

(2) 增强输入验证。防止网页篡改、信息泄露、木马植入等恶意网络攻击。

(3) 基于规则的保护和基于异常的保护。基于规则对应用进行全方面检测，并基于合法应用数据建立模型，以此为依据判断应用数据的异常。

(4) 状态管理。判断用户的访问次数，记录用户访问事件，检测异常事件，并在达到极限值时进行处理。

创新点：

(1) 采用 OpenResty 高性能服务端解决方案以及 Lua 语言进行开发，实现非阻塞的访问网络 I/O。在连接 MySQL、Redis 和发起 HTTP 请求时，工作进程需要支持事件驱动，用协程的方式让 CPU 资源更有效的去处理其他请求。

(2) 有完备的缓存机制。

(3) 利用规则库和数据模型协同检测。

六、毕设的基本思路

1. 收集相关资料，学习 OpenResty 以及 Nginx 开发技术；
2. 设计组件的总体架构，规划各个模块功能；
3. 实现各个模块的功能；
4. 整合并打包组件、测试功能。

七、论文写作提纲

第 1 章 绪论：介绍论文的选题背景和意义、论文的组织结构。

第 2 章 相关知识介绍：本章主要对实现该组件所用到的技术和框架进行介绍。

第 3 章 需求分析：本章主要对组件的总体需求等进行分析。

第 4 章 组件设计：本章主要介绍组件的整体架构以及问题解决方案。

第 5 章 组件实现：本章主要介绍组件核心功能代码的实现。

第 6 章 测试和实验：本章主要对组件相关功能测试结果进行展示。

第 7 章 总结与展望：对毕业设计内容做出总结。

八、主要参考文献

[1] 刘志宏,孙长国.基于 Web 访问日志的异常行为检测[J].计算机与网络,2015,41(13):62-64.

[2] 田润. 基于机器学习的参数注入式攻击检测方法研究[D]. 内蒙古:内蒙古农业大学,2019.

[3] 李曼玉. Web Service 中间件服务系统的设计与实现[D]. 北京:北京交通大学,2018.

[4] 陈鹏炜. 基于负载均衡与缓存技术的实名鉴权系统设计与实现[D]. 成都:电子科技大学,2017.

[5] 李嘉伟. 多维自适应 Web 异常检测系统研究与实现[D]. 北京:北京邮电大学,2017.

[6] 刘泽宇. 基于 Web 轨迹的应用层 DDoS 攻击检测[D]. 徐州:中国矿业大学,2017.

[7] 林旭. 基于 WEB 访问日志的异常检测技术研究[D]. 青岛:中国海洋大学,2015.

[8] 李浩杰. 基于 Web 日志的异常检测分析研究[D]. 西安:陕西师范大学,2015.

[9] 靳莹. 基于缓存技术的内容管理系统研究[D]. 吉林:吉林大学,2014.

[10] 罗浩然. Web 应用防火墙中流量处理模块的设计与实现[D]. 南京:南京大学,2019.

[11] 王晗. 基于容器技术的集成化测试系统的设计与实现[D]. 北京:北京邮电大学,2019.

[12] 赵正旭,申跃杰,左宗成.脚本语言 Lua 与 C++语言交互方法的研究[J].电脑知识与技术,2018,14(23):135-137.

[13] J Wang,L Xia. Abnormal behavior detection by causality analysis and sparse reconstruction[J]. Journal of Central South University,2017,24(12)

[14] B Kim,Kitae Park,T Kim,D Seo. Abnormal behavior detection algorithm of infra-structure using unfamiliarity index[P]. Smart Structures and Materials + Nondestructive Evaluation and Health Monitoring,2017

写作进度及具体时间安排	
起止日期	主要研究内容
2020.02.01-2020.03.01	实现基于规则的异常保护与状态管理功能
2020.03.02-2020.04.09	实现基于合法应用数据建立模型进行异常判断的功能
2020.04.10-2020.04.15	实现程序的整合以及打包

指导教师对开题报告的意见

指导教师签名：　　年　　月　　日

许昌学院本科毕业论文（设计）教师指导记录表

学　院	信息工程学院	指导教师	平源	职称	副教授
学生姓名	张释文	专　　业	网络工程	班级	2016 级本科 1 班
论文（设计）	基于 OpenResty 的 Web 异常访问行为检测组件设计				

初稿指导意见：

该论文初稿完成了对选题背景、研究现状、需求分析、概要设计、详细设计等内容的初步撰写，其中对于需求分析、概要设计、详细设计三部分的描述体现了该生对本课题清晰的认识和对技术路线的准确把握，行文逻辑基本合理、语句通顺，基本符合论文规范。但仍然存在较多不足，建议如下：

1. 论文摘要应表述的三部分逻辑还需优化，尽量通过个段落展示，其中涉及的技术、方法、属性（所设计系统或创新的属性）应该更加直接阐述；

2. 第 2 节“需求分析”应具体阐述需求分析方面的工作，与背景问题区分开；

3. 第 5 节“测试和实验”中的测试过于简略，至少应具备测试目的、测试方案、测试过程与结果（功能、性能等测试）；

4. 论文多处使用图例，但要正确区分和使用系统模型图、架构图、数据流图等。建议在需求分析部分从用户角度给出一个系统模型图，以增强对系统需求的表达能力；

5. 注意文中编号使用规则，具体见在初稿中的标注说明；

6. 论文中出现的英文、数字应使用 Times New Roman 格式，图片中的英文字符也使用 Times New Roman 格式；

7. 总结和展望应该更加具体一些。

指导教师签名：

年　　月　　日

二稿指导意见：

该论文二稿完成了对摘要、需求分析、概要设计、详细分析和测试实验等内容的修改与完善。相较于初稿，论文摘要应表达的逻辑得到改善，详细设计内容更加细致详尽，程序测试体现出对程序的测试更加充足。虽然有些改进， 但是不足之处仍然较多，建议如下：

1. 论文摘要关键词没有准确描述出所设计系统的特点，应选取准确描述所设计系统特点的词，并按照概念范围大小依次排列；
2. 论文摘要的表述仍然欠缺，在文中应有足够的理由证明方案的可行性；
3. 第 5 节的程序流程图应仔细切分为两个部分；
4. 文中编号使用格式有误，正文中编号应与标题编号进行区分；
5. 论文最后应有致谢页。

指导教师签名：

年　　月　　日

定稿前指导意见：

该论文定稿完成了对本课题研究、设计与实现的详细介绍与说明，并记录了所设计系统进行一系列程序测试的结果。论文整体逻辑合理、结构清晰，行文格式符合要求规范，摘要阐述逻辑清晰明确，设计思路与实现过程描述详细、逻辑表达完整，程序测试内容全面。虽然基本符合本科生论文定稿要求，但是仍存在些许不足，建议如下：

1. 部分图片不够清晰，需修改为清晰版本；

2. 第 2 章需求分析中，应该着重于从用户角度对需求进行分析，应该更详细说明，功能需求与非功能需求都要尽可能详细；

3. 总结与展望有部分冗余内容，建议删除。

指导教师签名：

年　　月　　日

许昌学院本科毕业论文（设计）中期检查表

<table>
<tr><td rowspan="4">学生填写内容</td><td>学　院</td><td>信息工程学院</td><td>学生姓名</td><td>张释文</td><td>专业</td><td>网络工程</td></tr>
<tr><td>论文（设计）题目</td><td colspan="5">基于 OpenResty 平台的 Web 异常行为检测组件设计</td></tr>
<tr><td>论文（设计）进展情况</td><td colspan="5">已完成组件各个功能模块设计与实现并完成论文初稿撰写</td></tr>
<tr><td colspan="6">论文（设计）撰写中存在的突出困难及解决办法
突出困难：机器学习模型误报率过高的情况下模型调优遇到困难。
解决方法：优化特征工程并采集质量更高的数据集进行调优。
学生签名：
年　　月　　日</td></tr>
<tr><td rowspan="8">指导教师填写内容</td><td colspan="3">检查评价内容</td><td colspan="2">实际工作状态</td></tr>
<tr><td colspan="3">与开题报告相比较，毕业论文(设计)的题目和内容有无调整</td><td colspan="2">无调整</td></tr>
<tr><td colspan="3">学生论文(设计)所取得的阶段性成果</td><td colspan="2">项目开发初步完成，论文初稿 70%</td></tr>
<tr><td colspan="3">学生的工作态度、出勤情况</td><td colspan="2">毕设认真、态度好，按时出勤，及时沟通进展</td></tr>
<tr><td colspan="3">指导教师对学生的指导情况(指导次数、方式)</td><td colspan="2">因疫情影响，以 QQ、电话、邮件指导为主，平均 1 周 1 次，已指导 8 次</td></tr>
<tr><td colspan="3">对能否按期完成毕业论文(设计)的评估</td><td colspan="2">应能按计划完成</td></tr>
<tr><td colspan="3">学生与指导教师有关毕业论文(设计)的原始材料是否保存齐全</td><td colspan="2">资料齐全</td></tr>
<tr><td colspan="5">指导教师对论文（设计）进展情况的意见：
该毕业设计题目为师生多次协商确定，并采用了较高的难度设置。目前，虽然整体进展较之预期略有滞后，但经过最近（4 月 2 日、4 月 10 日）两次对进展的深度沟通和方案改进，确保该设计应能如期完成。
指导教师签字：
年　　月　　日</td></tr>
<tr><td colspan="6">系（教研室）意见：
系（教研室）主任签字：
年　　月　　日</td></tr>
</table>

主管院长签名（盖章）：　　年　　月　　日

许昌学院本科毕业论文(设计)指导教师审阅意见表

学　院	信息工程学院	指导教师	平源	职称	副教授
学生姓名	张释文	专　　业	网络工程	班级	2016 级 本科 1 班
论文（设计）题目	基于 OpenResty 平台的 Web 异常行为 检测组件设计			字数	20148

该论文定稿完成了对本课题研究、设计与实现的详细介绍与说明，并记录了所设计系统进行一系列程序测试的结果。论文整体逻辑合理、结构清晰，行文格式符合要求规范，摘要阐述逻辑清晰明确，设计思路与实现过程描述详细、逻辑表达完整，程序测试内容全面。论文基本符合本科生论文定稿要求，但是仍存在些许不足，建议如下：

1. 论文中使用的图示，其图片和图名应保持在同一个页面；

2. 第 2 章需求分析中，应该着重于从用户角度对需求进行分析，应该更详细说明，功能需求与非功能需求都要尽可能详细；

3. 第 6 章“测试和实验”中检测指标以表格形式展示会更直观。

成　　绩	

是否同意答辩

指导教师签名：

年　　月　　日

许昌学院本科毕业论文(设计)评阅教师评阅意见表

<table>
<tr><td>评阅人姓名</td><td>马慧</td><td>单位</td><td>信息工程学院</td><td>专业</td><td>网络工程</td><td>职称</td><td>副教授</td></tr>
<tr><td>学生姓名</td><td>张释文</td><td>专业</td><td>网络工程</td><td>班级</td><td colspan="3">2016 级本科 1 班</td></tr>
<tr><td>论文题目</td><td colspan="4">基于 OpenResty 平台的 Web 异常行为检测组件设计</td><td>字数</td><td>20148</td></tr>
<tr><td colspan="8">该论文在选题上，符合专业培养目标毕业论文，难度适中、工作量较大。题目以解决实际问题为目的，提出将机器学习技术与正则匹配结合的方式进行 Web 异常访问行为的检测，具有一定的创新应用和实践指导价值。
论文在内容上，共分为七个章节，层次结构安排合理。中英文摘要表达逻辑清晰，对该课题介绍准确。需求分析、概要设计和详细设计内容完整，对设计思路和设计过程的具体细节描述详细。测试和实验方案合理，对所设计系统进行了多方面的测试。论文在格式上，基本符合学院要求规范，无任何明显错误。虽然基本符合要求，但是还存在些许不足，如下：
1. 论文需求分析还存在表达含义不够准确的问题。
2. 论文图例表达含义还不够明确，建议再次改进。
3. 论文参考文献没有详细表明页码。</td></tr>
<tr><td>成　　绩</td><td colspan="7"></td></tr>
<tr><td colspan="8">是否同意答辩

评阅教师签名：
年　　月　　日</td></tr>
</table>

许昌学院本科毕业论文(设计)答辩记录表

学生姓名	张释文	性别	男	专业	网络工程	班级	2016 级本科 1 班
指导教师姓名	平源	职称	副教授	答辩时间	2020.05.22	答辩地点	信工楼 311（线上）
题目	基于 OpenResty 平台的 Web 异常行为检测组件设计					字数	20148

问题 1：为什么要使用逻辑回归算法？

回答：使用逻辑回归算法做二分类模型进行检测，对正常行为的检测准确性较高，这样可以保证组件机器学习模块误报率降低，若是漏洞还可以通过后续的规则引擎进行检测。

问题 2：性能评测的标准是什么？各方面的硬件配置如何？

回答：这里使用万分制进行评分，评分越高性能表现越好。主要根据 CPU 的使用率、内存使用率得出评分。CPU 是双路八核十六线 CPU，内存 32G，带宽为 100M。

问题 3：性能测试的三张图分别代表什么？

回答：这里分别是三次检测结果的图表统计，放在这里做一个对比，最后一张图是选取的最坏情况下的性能评分。

问题 4：如果给某一个企业的网做入侵检测，能够抗过它的什么流量，自己测过了没有？离实际的产品有多远？

回答：对于中小型企业可以很好的进行防护，对于特大型网站还需要进行更专业的性能优化。我的这款组件可以抗衡目前的开源 WAF，但是如果与其他产品进行对比的话，我这里没有大量的真实网络环境数据，采集到的数据用来做组件的模型优化，使用我采集的数据进行检测对比有失公平性。因此，成为实习的产品还需要在真实的网络环境中进一步验证优化。

答辩小组秘书签名：

年　　月　　日

<table>
<tr><td rowspan="6">答辩分项成绩</td><td>构成</td><td>姓 名</td><td>职称</td><td colspan="2">系（教研室）</td></tr>
<tr><td>组长</td><td>马 慧</td><td>副教授</td><td colspan="2">网络工程系</td></tr>
<tr><td rowspan="3">答辩小组成员</td><td>张志立</td><td>教 授</td><td colspan="2">网络工程系</td></tr>
<tr><td>孙培岩</td><td>讲 师</td><td colspan="2">网络工程系</td></tr>
<tr><td></td><td></td><td colspan="2"></td></tr>
<tr><td colspan="2">开题报告成绩</td><td></td><td>预答辩成绩</td><td></td></tr>
<tr><td colspan="2">终期答辩成绩</td><td colspan="3"></td></tr>
<tr><td>答辩总评成绩</td><td colspan="5"></td></tr>
<tr><td></td><td colspan="5">计算方式：开题报告 10%，预答辩 30%，终期答辩成绩 60%</td></tr>
<tr><td rowspan="7">终期答辩评分标准</td><td colspan="4">评分项目</td><td>满分</td></tr>
<tr><td colspan="4">工作量</td><td>10</td></tr>
<tr><td colspan="4">技术水平</td><td>25</td></tr>
<tr><td colspan="4">研究成果基础理论与专业技术</td><td>20</td></tr>
<tr><td colspan="4">文字表达</td><td>5</td></tr>
<tr><td colspan="4">答辩效果</td><td>40</td></tr>
<tr><td colspan="4">合 计</td><td>100</td></tr>
<tr><td colspan="6">答辩小组意见

组长签字：
年 月 日</td></tr>
<tr><td colspan="6">学院意见

学院签章：
年 月 日</td></tr>
</table>