2020 年全国大学生信息安全竞赛

作品报告

作品名称：慧眼——基于客户端蜜罐和机器学习的风险网站检测系统

电子邮箱：824342218@qq.com

提交日期：2020 年 6 月 15 号

填写说明

1. 所有参赛项目必须为一个基本完整的设计。作品报告书旨在能够清晰准确地阐述（或图示）该参赛队的参赛项目（或方案）。

2. 作品报告采用A4纸撰写。除标题外，所有内容必需为宋体、小四号字、1.5倍行距。

3. 作品报告中各项目说明文字部分仅供参考，作品报告书撰写完毕后，请删除所有说明文字。(本页不删除)

4. 作品报告模板里已经列的内容仅供参考，作者可以在此基础上增加内容或对文档结构进行微调。

5. 为保证网评的公平、公正，作品报告中应避免出现作者所在学校、院系和指导教师等泄露身份的信息。

目录

摘要.....6

第一章 作品概述.....9

1.1 背景综述.....9

1.1.1 进行风险网站检测的必要性.....9

1.1.2 相关工作.....12

1.1.2.1 现有产品分析和不足.....12

1.1.2.2 相关问题解决方案.....15

1.2 功能详述.....16

1.2.1 进行风险网站检测.....17

1.2.2 检测结果展示.....18

1.2.3 用户申诉与举报.....18

1.3 作品特性.....19

1.4 可行性分析.....20

1.4.1 技术可行性分析.....20

1.4.1.1 客户端蜜罐.....20

1.4.1.2 自然语言处理技术.....21

1.4.1.3 Vue 等前端技术.....22

1.4.2 市场可行性分析.....22

1.4.2.1 对风险网站检测的需求.....22

1.4.2.2 对当前现有产品的研究.....23

1.4.2.3 系统性能和市场相结合.....23

1.5 作品特色.....24

1.5.1 设计客户端蜜罐辅助进行恶意软件分析.....24

1.5.2 确立了动态解析的 JS 反混淆机制.....24

1.5.3 基于预取的钓鱼网站检测系统.....25
1.5.4 基于自然语言处理技术的不良信息网站检测技术.....25
1.5.5 基于响应式网页设计的数据图表展示.....26
1.6 展望.....26
第二章 作品设计与实现.....27
2.1 系统架构.....27
2.1.1 系统架构分层视图.....27
2.1.1.2 展示层.....28
2.1.1.3 业务层.....28
2.1.1.4 数据层.....30
2.1.2 系统开发技术.....30
2.1.2.1 前端开发相关技术.....30
2.1.2.2 服务端开发相关技术.....31
2.1.2.4 自然语言处理相关技术.....31
2.1.2.5 机器学习相关技术.....33
2.1.2.6 病毒扫描相关技术.....35
2.1.3 系统运行环境.....36
2.1.3.1 浏览器.....36
2.1.3.2 服务器环境.....37
2.2 服务流程.....37
2.3 功能模块.....38
2.3.1 系统日志记录模块.....38
2.3.1.1 目的.....38
2.3.1.2 类设计.....39
2.3.2 用户管理模块.....39
2.3.2.1 目的.....39
2.3.2.2 模块设计.....40
2.3.3 用户申诉举报及管理员裁决模块.....40

3.2.3.1 目的 40
2.3.3.2 模块设计 41
2.3.4 邮件模块 41
2.3.4.1 目的 41
2.3.4.2 模块设计 42
2.3.5 恶意软件下载检测模块 42
2.3.5.1 目的 42
2.3.5.2 实现方法 43
2.3.6 钓鱼网站检测模块 44
2.3.6.1 目的 44
2.3.6.2 实现方法 44
2.3.6.3 理论分析 45
2.3.6.4 检测流程 45
2.3.6.5 检测特点 46
2.3.7 不良信息网站信息检测模块 46
2.3.7.1 目的 46
2.3.7.2 算法原理说明 46
2.3.7.3 判断过程 49
第三章 作品测试与分析 50
3.1 引言 50
3.1.1 编写目的 50
3.1.2 测试范围及方法 50
3.1.3 测试环境 51
3.1.4 系统可能风险 52
3.1.5 测试结束条件 52
3.2 系统功能测试过程 54
3.2.1 客户端功能测试 54
3.2.1.1 查询功能 54

3.2.1.2 用户注册 56
3.2.1.3 网站位置分布与危险等级显示 57
3.2.1.4 用户申诉 58
3.2.1.5 管理员登录后台 59
3.2.1.6 管理员裁决 60
3.2.1.7 产品信息安全测试 60
3.2.1.7.1 测试工具与方法 61
3.2.1.7.2 测试结果 61
3.2.2 服务端功能测试 62
3.2.2.1 生成管理员账户 62
3.2.2.2 删除数据库中的网站 63
3.3 对比测试 64
3.3.1 测试说明 64
3.3.2 总体检出率对比测试与分析 65
3.2.3 恶意软件下载网站检出对比测试与分析 65
3.2.4 钓鱼网站检出对比测试与分析 66
3.2.5 色情网站检出对比测试与分析 67
3.2.6 博彩赌博网站检出对比测试与分析 68
3.2.7 假阳性率对比测试与分析 68
3.4 测试总结与改进空间 69
3.4.1 总结 69
3.4.2 改进空间 70
第四章 创新性说明 71
4.1 系统性创新 71
4.2 模块内创新 72
4.2.1 设计客户端蜜罐辅助进行恶意软件分析 72
4.2.2 确立了动态解析的 JS 反混淆机制 72

4.2.3 基于预取的钓鱼网站检测系统.....73
4.2.4 基于自然语言处理技术的不良信息网站检测技术.....73
4.2.5 基于响应式网页设计的数据图表展示.....73
第五章 总结.....74
5.1 系统设计与开发.....74
参考文献.....76

摘要

近几年，我国云计算、大数据、物联网、工业互联网、人工智能等新技术新应用大规模发展，网络安全风险融合叠加并快速演变。互联网技术应用不断模糊物理世界和虚拟世界界限，对整个经济社会发展的融合、渗透、驱动作用日益明显，带来的风险挑战也不断增大，网络空间威胁和风险日益增多。网络黑产活动专业化、自动化程度不断提升，技术对抗更加激烈。便捷的网络服务吸引了网络攻击者们通过钓鱼网站，色情网站和恶意软件推广等方式非法牟利。尽管这些不法活动的目的和手段各不相同，但它们都需要不知情的用户访问攻击者提供的网页地址以达到攻击目的。

尽管目前学界对于风险网站的定义尚没有一个确定的说法，但在我们这个产品中，我们指的是**广义的风险网站:即不单单指传统的恶意软件下载，挂马，网站伪冒钓鱼这种恶意网站；还包括色情网站，博彩网站这种包含不良内容的网站，我们都统称为风险网站。**具体我们把风险网站细分为两大类四小类，具体的如图 1 中所示。

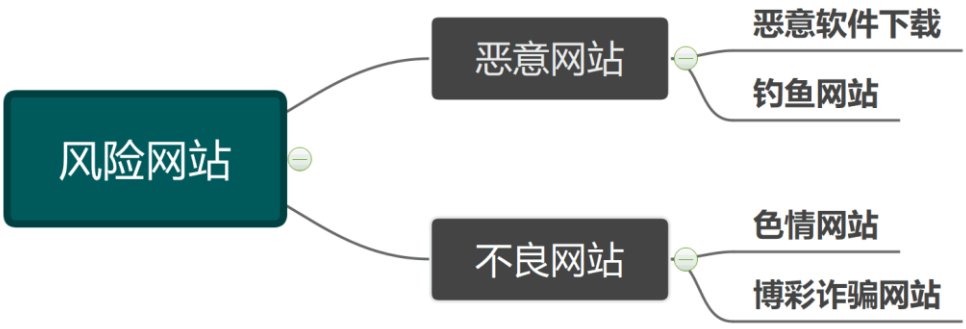

图 1

对于恶意网站，据《2019 年中国网络安全报告》，在过去的 2019 一年中，仅瑞星“云安全”系统就截获病毒样本总量 1.03 亿个，病毒感染次数 4.38 亿次，

病毒总体数量比 2018 年同期上涨 32.69%。给我国人民财产造成了巨大的损失。截至撰文时，在最新的《2020 年 3 月互联网安全威胁报告》中体现出：网络病毒活动情况方面,境内感染网络病毒的终端数为近 151 万余个,较上月增长 13.7%。网站安全方面境内被篡改网站数量为 26,029 个，较上月增长 41.6%，同时，在 CNCERT 处理的事件中，在 9472 件事件报告中，排名前三位的安全事件分别是恶意程序、漏洞、网页仿冒。

另一方面,在不良内容网页方面，情况也不容乐观。据新华社的一份报告，淫秽色情信息是网民最容易接触到的互联网不良信息,相关举报占到了举报总量的 76.3%。还有 10.7%的举报信息为衣着暴露或行为不雅的不良信息。这些信息虽然还不构成淫秽色情信息的性质，但持久下去也会对社会风气产生不良影响。此外，赌博诈骗等博彩信息占到了 11.2%，值得注意的是，在这 11.2%的博彩信息中，混杂在淫秽色情信息中进行传播的。如虚假的同城交友、视频交友，色情直播平台等，其最终目的都是诈骗。**从报告可以看出，色情网站和博彩网站占到了所有不良信息网站的 98.2%，这也将是网站信息识别的重点方向。**

在上面的阐述中可以看出，各种类型的风险网站无论是从数目还是范围有着愈演愈烈的趋势，有一款产品来对 URL 进行正确且全面的识别已经是迫在眉睫。通过对现有检测系统的研究，我们发现他们可以实现基本的检测功能，但是仍然存在**检测类型不足、过于依赖黑名单或已有签名**等各方面的缺点。

风险网站在过去形式趋同，容易被轻松识别，但到现在，风险网站的特征和界面也在不断变化，检测的难度随之上升。一般的**静态检测方案不能很好地检测日益变化的风险网站**。

鉴于以上的问题，我们设计并完成了一款**基于客户端蜜罐和机器学习的风险网站检测系统**,可以针对性的对目前市面上系统的相关不足提出针对性的改进。该系统实现**了黑名单法+蜜罐检测法+机器学习三种方法的多维度结合**,并且**为不同类型的风险网站设计了不同的识别方案**。从而大大提高了检测的准确率和针对性。根据我们的产品特性，我们将它命名为“慧眼”。

与传统的服务端蜜罐不同，客户端蜜罐并不是守株待兔般的等待攻击者前来入侵。而是客户端本身去**主动访问**需要检测的风险网站。通过对网页上的代码进行**动态解析检测**，检测这一过程中是否有新进程产生，系统文件修改，注册表修改从而来判别网页是否含有恶意代码。

我们在对恶意软件下载类网站的检测上创新性的引入了**客户端蜜罐系统**来辅助检测。这一举措使得网页代码能够在一个类似沙箱的环境进行运行，同时动态执行的 JS 代码能够使得系统更加高效，准确的识别出恶意的 JS 代码，给系统带来了普通算法分析 JS 代码所无法具有的准确性。

在钓鱼网站检测和不良信息检测方面，系统对这两种类型的风险网站采取了不同的检测策略。对于钓鱼网站，我们发现当前市面上使用的基于 HTML 文本的检测，基于视觉的检测都很容易因为 HTML 代码的灵活性和没有及时更新特征库导致检测率下降。于是我们采用了**基于预取的钓鱼网站检测方法**，利用钓鱼网站和正常网站拓扑结构的不同进行识别，这种识别方法不局限于单个网页而是着眼于全局的拓扑结构，使得检测的准确率大大提升。

在不良信息网站检测方面，我们不只着眼于市面上常用的基于文本的 TF-IDF 算法对内容进行识别，同时还利用了不良信息网站相较于正常网站，站点具有域名上分隔符较多，字符转化次数多，分隔符间长度较低等特点，于是我们以此特点利用 SVM 构造分类器来实现对不良信息网站的检测。我们通过这两种方法的结合，成功的使系统对不良信息网站的检出率达到了 95%以上。

与市面上的其他系统相比，慧眼系统创造性的**对黑名单法、客户端蜜罐法、机器学习法三种方法进行了有机结合**，使得系统对风险网站的检出率大大超过了市面上其他各种系统。并且由于针对性的对不同类型的网站设计了不同的算法进行分析，使得系统在分项检出率、假阳性率上的表现也明显好于其他产品。我们的作品可以作为公共安全领域的一个基础业务，提供给用户较为精确的风险网站检测服务。

第一章 作品概述

1.1 背景综述

1.1.1 进行风险网站检测的必要性

截至 2020 年 3 月，我国网民规模为 9.04 亿，成为了世界上网络规模最大的国家，互联网普及率达 64.5%，庞大的网民构成了中国蓬勃发展的消费市场，也为数字经济发展打下了坚实的用户基础。一方面互联网的蓬勃发展为人们的日常生活创造了巨大的便利条件，但是在繁荣的表面下，底下潜藏的威胁从未离我们远去，反而有愈演愈烈之势。

在恶意网站方面，网络攻击者们使用恶意程序来窃取信息，入侵用户主机；制作钓鱼网站欺骗用户，达到窃取用户提交的银行账号、密码等私密信息的效果。在不良信息方面，攻击者往往通过制作色情网站来进行牟利，利用色情网站来对人们的精神进行侵蚀；通过制作赌博诈骗等类型的博彩网站来达到在网上赌博进行牟利的目的。

我们先从数据来看一下各种风险网站的危害和严重性。恶意网站方面，2019 年，我国境内感染计算机恶意程序的主机数量达到了惊人的 582 万台。其中通过恶意网站下载感染的主机数量超过一半；而在钓鱼网址欺诈方面，情况同样严峻，在 2019 年，仅 CNCERT 一个平台，就检测到了 8.5 万个针对我国境内网站的网站伪冒攻击，比 2018 年同比增长 59.7%。

不良信息网站方面，情况同样不容乐观。仅 2020 年 5 月一个月，全国各级网络举报部门受理举报 1519.9 万件，环比增长 4.2%、同比增长 21.5%。全国主要网站受理举报 1374.0 万件，环比增长 6.3%，同比增长 51.3%。可以看到在各类指标上，全国不良信息网站数量都呈增长势头。不难想象，如果不加以控制，各类不良信息网站容易趋于泛滥。

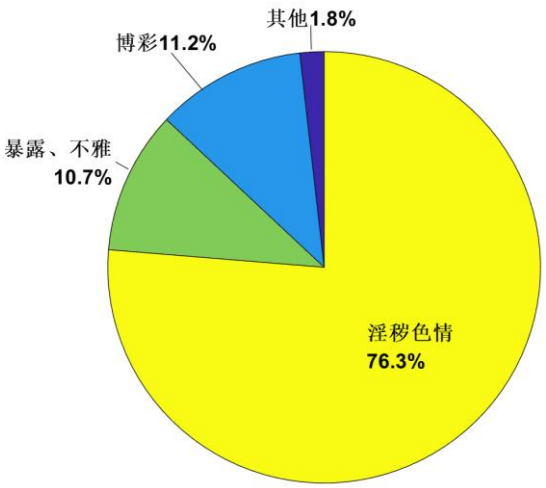

图 1.1.1(a) 2019 年各类不良信息网站占比统计

具体到各种类型的不良信息网站方面，从图表 1.1.1(a)可以看到，在所有不良信息网站中，色情和诈骗博彩类占到了 98%以上，成为了危害最大的不良信息网站类型。

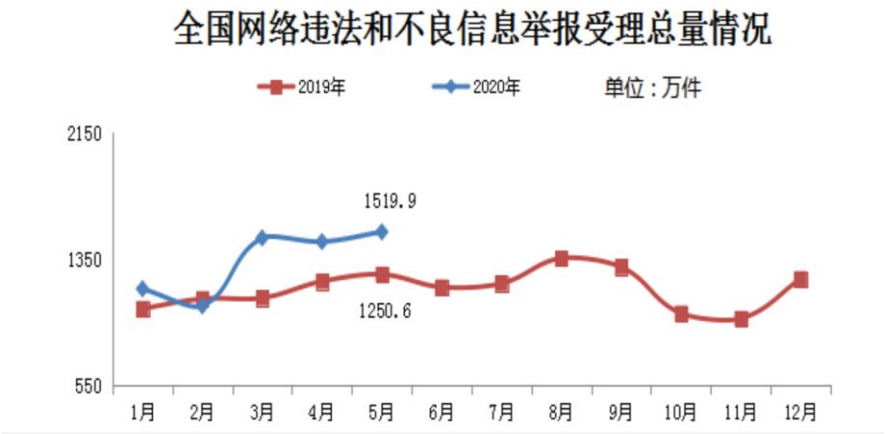

图 1.1.1(b) 2019 和 2010 年全国违法和不良信息举报受理情况

从图表 1.1.1(b)可以看出，**当前我国网络风险网站态势十分严峻，恶意网站和不良信息网站这两大风险网站类型均有泛滥日趋泛滥的风险，给我国互联网生态造成了相当严重的破坏。**

各种各样的风险网站还会给国家和人民带来巨大的财产损失。相关的新闻在我们生活中比比皆是；例如最近的净网行动，北京市公安局网安总队会同海淀公安分局，打掉一个为境外电信网络诈骗人员提供钓鱼网站的犯罪团伙。该团伙协

助诈骗分子作案 80 余起，涉案金额约 300 万元人民币[4]。犯罪团伙往往瞄准中老年人或青少年等防范意识薄弱的团体，通过钓鱼网页等手段盗取受害人钱财，**给人民群众的财产安全带来了重大的安全隐患。**

风险网站还会危害人民的身心健康。这当中，问题最严重，最突出的就是青少年浏览色情淫秽网站。青少年进入青春期，性发育开始成熟，性意识开始出现，充满了对“性”的好奇、幻想和冲动。在这个阶段，他们愿意谈一些性问题，开始关注异性，同时也很想知道性关系到底是什么。但是目前由于社会、家庭、学校对性教育认识的不充分，孩子们对性知识的获取渠道不通畅，对性问题的辨别和认识能力不够，这促使他们利用别的途径获得信息。现在很多青少年性犯罪的产生，跟网络色情文化的冲击是分不开的，色情文化对心理冲动起到一种恶性的催化作用。使得青少年的心理萌动、冲动被激活，无法自抑，最后发展到寻求生理发泄的对象，从而走上犯罪道路。 例如 2014 年安徽省蜀山区某少年晓林在不到两月内强奸 6 名女孩，事后追究起因就是他在网吧里认识一些大哥哥。这些大哥哥手把手教晓林浏览色情网站，这让晓林迅速扭曲地“成熟”起来[5]。从上面这个实例可以看出，**风险网站会给青少年带来精神上的巨大危害，让青少年空耗时间在虚拟网站中，压力长时间得不到排遣还容进行相关犯罪活动，另一方面，色情网站往往和诈骗网站结合在一起，容易让青少年在不知情的情况下误点，进一步带来财产上的损失。**

这些由风险网站引发的各类事件，让我们感受到风险网站的巨大破坏力，风险网站不只会造成人民群众财产的损失，更会污染人们的精神世界，扰乱人们的正常生活，严重的话还好威胁到社会的稳定。**同时我们也意识到对风险网站进行有效检测的必要性，只有高效准确的检测出风险网站，才方便于有关部门进行及时准确的管控。**

加强对风险网站的及时监测、正确的筛选出风险网站并进行及时的管控，对维护社会稳定、促进国家发展具有重要的现实意义，也是创建和谐社会的应有内涵。而如何提高风险网站检测的准确性、全面性、及时性，是当前社会环境下信息监管部门首先需要解决的问题。

1.1.2 相关工作

1.1.2.1 现有产品分析和不足

对于风险网站检测，目前市面上已经存在多种在线检测系统，这些风险网站检测系统工作基本流程如下：

1. 从各个渠道收集待检测的网站域名，如搜索引擎、社交软件、新闻评论。保存至样本数据库。
2. 检测引擎调用利用浏览器引擎，处理样本网站,获取页面渲染后的有效内容。
3. 调用检测算法对页面内容进行分析检测。
4. 统计检测结果，生成详细的检测报文，包括网页有无不健康内容、恶意软件、盗号风险、木马。
5. 将检测报文输出至第三方的拦截系统、关停服务提供商、加入黑名单等，遏制“钓鱼攻击”的发生。

目前国内各大互联网公司几乎都提供网页安全检测服务，如：百度网址安全检测、腾讯网址安全检测中心、360 网址安全查询。国外出名的网页安全检测产品包括 virustotal、URLhaus、joesandbox。通过对以上这些风险网站检测系统的研究分析,我们发现,网页安全检测方法主要分为三大类:基于黑名单机制类、实时运行分析类、机器学习分析类，它们有各种优点和不足之处。我们选取几大著名的分析引擎进行了着重分析。

• JoeSandbox

JoeSandbox 是运行时分析的代表,可用于检测、分析、保护 Windows, Android, Mac OS, Linux, 和 ios 系统中的可疑行为。用户可以直接上传文件或者发送下载链接，通过设置目标操作系统、浏览器版本、Java 版本和 Flash 版本自定义沙箱，然后在沙箱内执行上传文件的深度分析。

采用沙箱机制使得恶意程序精心设计的外壳失去了作用，程序运行时，任何不正常的系统调用、读写都将被监控。如果检测到程序有恶意行为，就会发出警

报，告诉用户这个网站是恶意网站，应该避免访问。

正是因为这种让威胁被制止在沙箱内的方法，这类网页安全检测系统有着很高的准确率，而且能够识别变形后的、新出现的、未被发掘的恶意网站，不依赖恶意网址数据库。然而引入沙箱机制是有代价的，运行沙箱需要**占用服务器的大量资源**，无法同时为大量用户提供网页安全检测。相较于查询类，**沙箱检测的耗时也会更久**，当网页内容异常丰富时这个问题会更加严重。

沙箱类网页安全检测系统还有一个致命缺陷，它无法用于鉴别网页的真伪，也无法分析网页是否存在不健康内容，毕竟，沙箱类网页安全检测系统的设计初衷就是为了检测出网页中的病毒、木马，至于网页本身的内容是什么，它并没有分析。

·腾讯网址安全检测中心

腾讯网址安全检测是基于黑名单机制类安全检测系统的代表，它的功能相对与沙箱类安全检测系统更为简单，相当于维护着一个风险网址的黑名单，任何人都可以通报自己在网页浏览中碰到的风险网站，只需要填写网站的 url，描述风险网页的类型，等待审核通过后，就可以将其加入风险网址的黑名单。**虽说腾讯网址检测中心能够对明显的文字关键字及图片进行一定程度的识别，但是主要还是依赖于腾讯产品的巨大用户进行举报识别后所得。**

这种检测方式就像是排雷，只要有一个人发现某风险网站，通报后进入黑名单，后来的其他人访问这个网站时就会被警告，从而避免了更多人受到此风险网站的危害。

相比沙箱类安全检测系统这种检测方式的速度快很多，毕竟就只需要完成一个类似查找表的工作，也不需要多少系统资源。所以，黑名单机制类安全检测系统是一种低成本条件下能获得不错效果的方法。

不过，黑名单机制类安全检测系统的效果几乎完全取决于用户提交的风险网站样本，如果只有很少用户提交，风险网站黑名单就很不完整，检测的准确率就很低。而且，风险网站的域名经常变更，在黑名单上的网页只要改变一下 url 就可以避开检测了。对于这种情况，黑名单机制几乎形同虚设。

同时，腾讯检测平台对于如钓鱼,色情类型的页面尚具有较高的识别率，但对于恶意文件下载这种类型的恶意网站识别率则较为低下。而且，在实际使用中，

发现腾讯网址安全中心可能维护精力较少，出现了不少 Bug，例如当输入的检测网站是 IP 形式时，将长时间没有返回结果，这对于实际进行 API 接口调用是一个相当棘手的问题。

·Yalih

Yalih(Yet Another Low Interaction Honeyclient) 堪称综合了黑名单机制类安全检测系统和沙箱类网页安全检测系统的优点，因为它将黑名单机制和沙箱机制的结合，形成了独特的蜜罐式的网页安全检测系统。

所谓蜜罐，就是用于检测、分析网页中或者浏览器插件中的恶意脚本，Javascript 是几乎所有浏览器都支持的语言，它可以给网站开发人员带来高响应高交互性的体验。然而它还给攻击者提供了跨平台攻击的机会，不论什么浏览器、什么操作系统、什么硬件架构。只要使用了 Javascript 脚本语言，漏洞利用都是相通的。

而 Yalih 对 Javascript 恶意脚本有很好的甄别能力，这得益于其采用了多种拥有反混淆、正则化功能的签名检测引擎，它还配备有一个可以设置 Cookie、重定向并且能够模仿主流浏览器 UA 的虚拟浏览器，防止恶意网站采用跟踪伪装技术，向蜜罐发送与正常用户不同的内容，可以同时达到较低的误报率并大大减少了扫描时间。

同时 Yalih 还集成了恶意文件签名机制，这是一种黑名单机制的改进，它引入第三方的恶意网址库，根据网站的签名先初步判断网页是否为恶意网站。

看似完美的 Yalih 也还是有不足之处，**Yalih 和沙箱类的产品类似，无法对网页内容进行识别,对钓鱼，色情，欺诈博彩等类型的恶意网站无能为力。**

基于对现状的分析，我们总结出现有网页安全检测系统所存在的三个主要问题

1.**对混淆的解构能力不强**，很多风险网站对 JS 代码进行一定程度的混淆，已有产品的检出率就会大大下降。

2.**大多数产品还是太过依赖于黑名单及已知文件的签名**，对风险网站的识别仍旧处在守株待兔的方法上，对机器学习技术的采用率不高。而采取了机器学习技术的系统，往往也并没有针对性的对不同类型的风险网站提出不同的检测算法。

3.**无法有效的对风险网站进行全方位的分类**，提供针对性的预警。如上面提

到的多种产品或在某种类型的风险网站检测上表现得不错，但都在某种程度上存在或多或少的“偏科”，也就是说，鉴于网络安全中“**木桶效应**”的致命性，社会急需要一种产品能够对全类型的风险网站进行**高效，精确，全方面**的扫描。

1.1.2.2 相关问题解决方案

针对上文所说当前系统存在的几个问题,我们设计出了"慧眼——基于客户端蜜罐和机器学习的风险网站识别系统".从以下几方面解决现有问题:

- **针对第一个问题，我们确立了动态执行的 JS 检测思路**

我们经过研究发现，当前 JS 混淆机制可以通过多种方法来进行混淆，如:base64,base95,甚至有些还采用了复杂的加密算法如(Feinstein and Peck, 2007, Heyman, 2007, Howard, 2010, Nazario, 2009)来进行相关的混淆。如果用算法进行解混淆分析，难免无法覆盖所有类型的混淆方法。

针对这种情况，我们的思路是先用特征检测的方法从网页中针对性的提取出.JS 文件(有的恶意网站会把 JS 直接嵌入到 html 代码中来逃避检测)。接下来的处理策略是先 Python 的 JSbeautify 模块进行一定程度的解混淆，再用 Rhino 引擎对提取出来的 JS 代码进行动态执行，监控此过程的内存调用、文件下载情况，以此来判定 JS 代码是否为恶意代码。通过这种动态执行的 JS 检测方法，我们可以大大提高恶意 JS 代码的检测准确率。

- **针对第二个问题，引入机器学习技术进行主动防御**

这里我们在不良信息网站检测和钓鱼网站部分都有着不少创新。具体主要体现在以下两个方面：

在不良信息检测方面，我们**采用了基于站点域名的识别+基于网页信息的识别。**我们一方面我们也发现不良信息网站的站点比起正常网站，具有域名上分隔符较多，字符转化次数多，分隔符间长度较低等特点，于是我们以此特点利用 SVM 构造分类器来实现对不良信息网站的检测；另一方面利用 TF-IDF 等文本聚类等相关技术手段提取网页中的关键信息，获取网页中的关键词，并引入 **SVM 向量机技术**，生成针对不良网站信息的分类器。利用分类器对来实现对网页信息的检测；通过采用基于站点的识别+基于文本的识别，我们实现了对不良信息网站的精准识别。

对于钓鱼网站，则不方便再利用上面的方法了，因为钓鱼网站可以实现在样式和结构和原网站的完全仿真,此时使用 TF-IDF 技术分析效果并不好。我们经过研究发现，钓鱼网站相较于正常网站，拓扑结构简单许多。这是由于钓鱼网站不用处理正常网站那么多的业务，由不法分子模仿正常网站复刻而成。于是我们基于此**采用了基于预取的钓鱼网站检测法**，通过分析钓鱼网站的拓扑模型和正常网页的拓扑模型进行对比进行训练，从而实现了较高的检出率。

- **针对第三个问题,我们构建了“三位一体”的检测体系**

针对市面上缺乏既能对恶意木马进行识别，又能对网站内容进行有效识别的风险网站识别系统，我们同时构建了黑名单法+客户端蜜罐法+机器学习识别法构建了三位一体的检测体系，以黑名单法为最底层可以获得较高的检测速度(如果黑名单里面有要检测的网页)。

蜜罐法可以**对恶意 JS 代码和恶意可执行文件进行动态分析**，实现了对恶意文件的高检出率和低假阳性率。**用不同的机器学习的算法来分析**网页内容，可以实现对不同种类网页内容的正确分析，采用的**自然语言分析**这一思路也使得正确率大大提高。

1.2 功能详述

“慧眼——基于客户端蜜罐和机器学习的风险网站识别系统”结合了客户端蜜罐技术和特有的网页内容感知能力，可以对用户输入的网站进行全面的网页内容检测。同时，还可以获取用户输入网站的 IP，经纬度，所在地区在世界地图上的位置，直观化，动态化的展示需检测网站的相关信息。可以帮助用户更清楚的获得相关结果。对于可能存在的误报问题，系统也提供了误报反馈功能让用户可以及时的向反馈结果以供管理员裁决。

1.2.1 进行风险网站检测

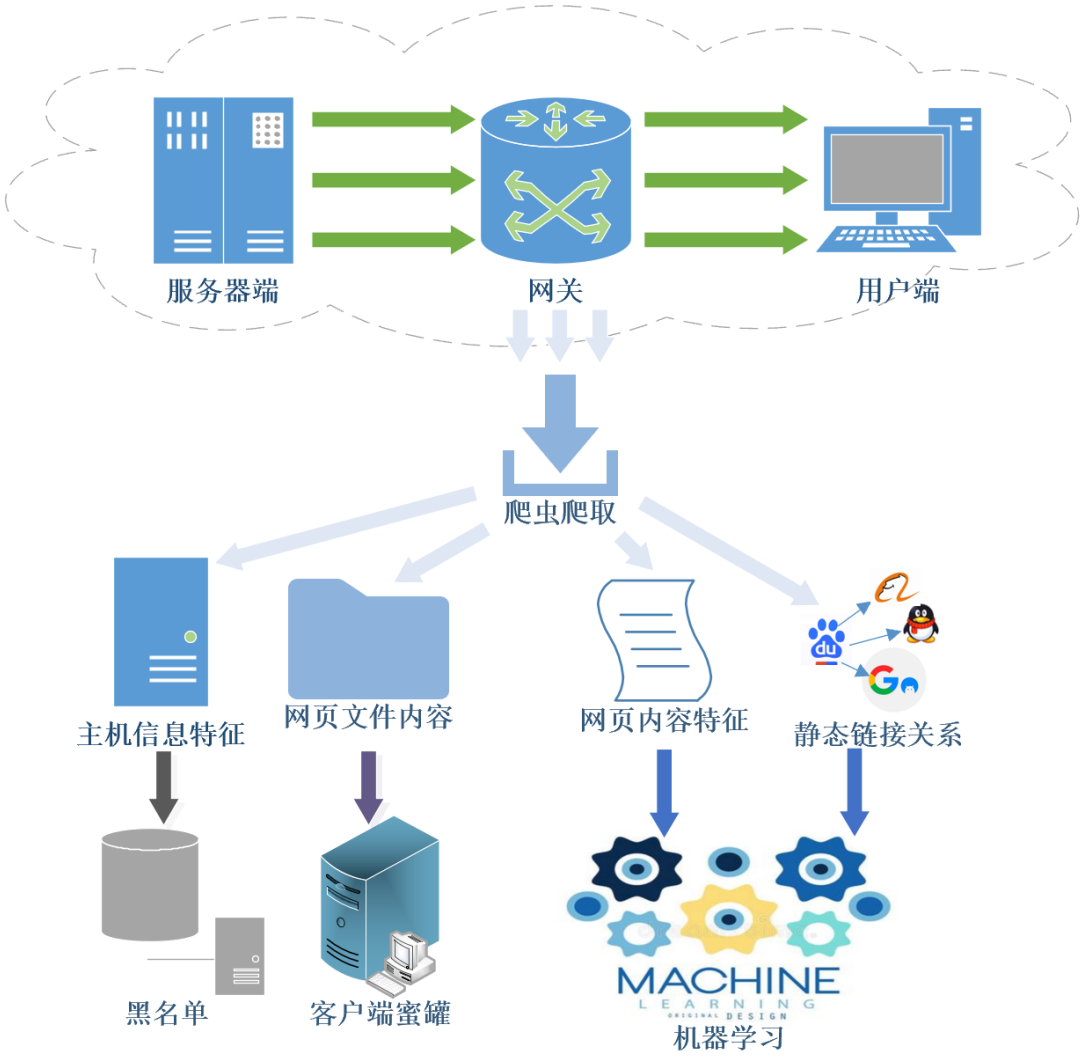

图 1.2.1 系统核心功能原理演示图

如图 1.2.1 所示，系统可以使用三大模块对需要检测的网站进行全面检测。通过获取特征和网页文件，在两个联网数据库中进行**黑名单检索**；网页文件则送入**客户端蜜罐**进行解混淆和动态检测；网页特征送入**机器学习模块**分别使用 TF-IDF 文本聚类技术，SVM 算法进行不良信息检测，同时使用基于预取的检测方法进行钓鱼网站识别。再返回检测结果。

1.2.2 检测结果展示

系统在分析网站后通过 Web 界面的形式将分析结果呈现给管理员，这样可以保证管理员在任何有网络的环境下均可以查看报表。报表将采用响应式网页设计(Responsive web design)，可使网站在多种浏览设备（从桌面电脑显示器到移动电话或其他移动产品设备）上阅读和导航，同时减少缩放、平移和滚动。

- **网站基本信息展示**

  网站基本信息展示返回网站的 IP，所在国家、省(州)、市、经纬度等网站基本信息。同时在一个可视化的世界地图上展示出网页所在地。给予用户最直观的信息展示。

- **检测结果信息展示**

  检测结果信息展示返回给用户总的检测结果还有三大模块对风险网站的检测结果，同时在此页面用户还可以根据检测结果选择是否进行申诉，以来协助系统消除假阳性和假阴性的情况。

- **风险网站地理位置分布图**

  风险网站地理位置分布图以节点模式把风险网站展示在地图中，并且其节点半径代表所在地区风险网站数目的多少。具体的区域范围用户可以根据自己的需求选择世界，国家，省(州)这三个尺度查看相关的风险网站数量。并且当用户把光标移动到具体节点上时，可以展示出改节点包含的风险网站的具体数量和各种类型风险网站的占比和数量。此图在一定程度上表现出了风险网站的地域分布规律（数量，密集程度等）。

1.2.3 用户申诉与举报

在实际应用中，所有风险网站系统都难免会出现检测的假阳性、假阴性事件。为此，我们特定引入了用户申诉与举报模块，并通过后端的邮件模块自动发送邮件给用户系统管理员的处理结果。

1.3 作品特性

本作品采用的蜜罐技术，TF-IDF 文本聚类技术，基于预取的机器学习技术均已有相关的论文进行描述，这在一定程度上保证了作品的可行性和可靠性。作品针对不同的网站类型制定了不同的检测策略，这使得系统检测的准确性大大提升。最终系统的报表呈现形式也易于用户操作，具有较高的易用性。该系统为用户提供了相关的 API，便于移植到多个平台和接口使用。

- **实用性**

随着近年来互联网的迅速发展,各种针对互联网用户的攻击层出不穷。不仅仅恶意网站给人们带来了巨大的财产损失及系统文件破坏，包含不良内容的网站也会给人造成精神上和心理上的不良影响，博彩类型的网站更是会让人陷入赌博的风险中、会给人民群众造成巨大的经济损失。鉴于以上种种情况，开发一种能够正确，高效识别这些类型的风险网站检测系统已经是迫在眉睫。因此，我们认为我们这款产品贴合了时代的痛点，有着极大的实用性。

- **可靠性**

系统采用时下较为成熟的技术，如客户端蜜罐、SVM 算法、TF-IDF 文本聚类技术、HTML5，在一定程度上避免了因技术不完善导致的错误。同时，服务器后端配置不使用性能较低的 Flask 自带 webserver 服务器，而采用 nginx + flask + uwsgi 进行配置，有效的提高了系统在高并发条件下的执行性。

- **易用性**

系统最终生成的数据报表将以 Web 界面的形式呈现给用户，采用 HTML5、JavaScript、响应式网页设计等技术则会带来更好的交互性，提高用户的使用体验，有较高的易理解性、易操作性。

- **灵活性**

系统可以通过多种方式进行调用，既可以通过 Web 页面进行访问使用，也可以 API 接口进行调用，方便灵活的接入各种系统中进行综合查询，实现了较高的使用灵活性。

- **高效性**

系统在构造时充分考虑了速度这一需求，充分运用多线程技术提高运行速度。同时在考虑算法和构造时也充分考虑了这一因素，例如在设计蜜罐时，没有选择采用拟合度更高但是会大大降低可靠性和效率的高交互蜜罐，而是选择在低交互蜜罐的基础上设计优秀的解混淆算法来提高准确性，从而同时满足了效率和精确性的要求。

- **高度集成性**

本系统集分析恶意网站和不良内容网站两大功能，基本上涵盖了所有风险网站类型的检测。同时，与其他系统不同的是，本系统集成了黑名单检测法+蜜罐检测法+机器学习分析法三大分析方法，这是市面上其他产品所不具有的，因此，本产品具有了极高的集成性，同时使用此三种方法进行分析，大大提高了产品检测的正确率。

1.4 可行性分析

1.4.1 技术可行性分析

1.4.1.1 客户端蜜罐

蜜罐（honeypot）本质上是一种用来发现攻击工具、攻击策略与攻击者攻击动机的知名技术，可以侦测或抵御未经授权操作或者是黑客攻击的陷阱，因原理类似诱捕昆虫的蜜罐而得名。

图 1.4.1.1 传统蜜罐技术示意图

生活中我们较常接触到的是服务端蜜罐，即部署在服务器上等待攻击者攻击，再来对攻击者的行为进行检测。而我们项目采用的是客户端蜜罐（Client Honeypot），其是模拟客户端去对可能存在恶意软件下载的网页进行主动模拟浏览，通过下载里面的文件或 JS 代码进行解混淆和动态解析，检测文件是否为木马文件。

本系统搭建了一个客户端蜜罐，使用 Python 的 mechanize 模块去模拟浏览器对网站的访问，解析出内嵌的 JS 代码和可执行文件，在客户端蜜罐中进行动态解析与反混淆，再配合 ClamAV 引擎对文件进行分析，从而达到检测恶意文件的目的。

1.4.1.2 自然语言处理技术

目前自然语言处理技术较为成熟，其中包含了如结巴中文分词、TF-IDF 算法、余弦相似性算法等等算法技术。系统使用自然语言处理等相关文本分析技术处理文本信息，通过使用现有比较成熟的网络爬虫技术，从色情网站，博彩网站上抓取数据，再利用 TF-IDF 等现有成熟的文本聚类技术对抓取到的数据（色情关键词，博彩关键词）进行词频分析、关键词提取等工作，参考「中央科学院现代汉语平衡语料库语料库」的相关格式，最终建立我们的「色情、博彩网站语料库」用以辅助检测。

1.4.1.3 Vue 等前端技术

Vue.JS 是一套构建用户界面的渐进式框架。与其他重量级框架不同的是，Vue 采用自底向上增量开发的设计。Vue 的核心库只关注视图层，并且非常容易学习，非常容易与其它库或已有项目整合。另一方面，Vue 完全有能力驱动采用单文件组件和 Vue 生态系统支持的库开发的复杂单页应用。比起其他重量级框架来，Vue 具有易用性，灵活性，高性能性等特点，考虑到上面种种因素，我们采用了 Vue 框架来开发我们的系统前端。

此外，我们还使用了 Chart.JS 等 JavaScript 图表库及 HTML5 等技术生成网站分布地图。这些图表能提供给使用者详细直观的交互体验。

1.4.2 市场可行性分析

1.4.2.1 对风险网站检测的需求

随着国家信息化发展战略的实施，我国网络基础设施建设取得了巨大成就。截至 2020 年 3 月，我国的互联网普及率达到 64.5%，成为世界上拥有最多网民的国家。网络的普及导致在线交易的增加，随之而来的网路诈骗行为也变得猖狂。

根据赛门铁克公司的报告，平均每 1126 个网站就有一个风险网站，而平均每个社交网络中，就存在 3378 个钓鱼网站，这些风险网站中存在着各式各样的欺诈行为，包括出售虚假商品、制作钓鱼网站、传播木马和病毒等，对用户的财产及信息安全造成了巨大的威胁。

为了避免用户的财产收到威胁、提高用户账户的安全性，识别风险网站是急需解决的一个问题。

同时，互联网的信息量日益增长，不良信息如：色情、血腥、诈骗、博彩也随之泛滥，在消耗大量网络资源的同时，也不利于社会风气建设，影响未成年人的身形健康，甚至诱导其走上违法犯罪的道路，危害社会的和谐稳定。因此，需要有一个可以及时，准确的风险网站检测系统来对各种各样的风险网站进行检测，从而辅助有关部门进行相关的查处。

1.4.2.2 对当前现有产品的研究

上文提到了现在虽然存在一些风险网站检测系统，它们也在一定程度上满足了风险网站的检测需求但它们仍然存在很大的不足。市面上现有的系统（如腾讯网址安全中心，Virustotal，JoeSandbox 等）普遍存在过度依赖黑名单，针对类型太少，解混淆能力不足等以及只依赖于同一模型，没有针对不同类型的风险网站设计出不同的检测思路等问题，导致系统的存检出率低、误报率高等问题。诸如此类问题，都是现有风险网站检测系统所没有办法解决的。

1.4.2.3 系统性能和市场相结合

检测的准确性固然重要，但是检测的速度同样不容忽视。我们团队通过问卷调查，调查了 96 人对于风险网站扫描时间的最大接受程度。如图 1.4.2.3 所示：

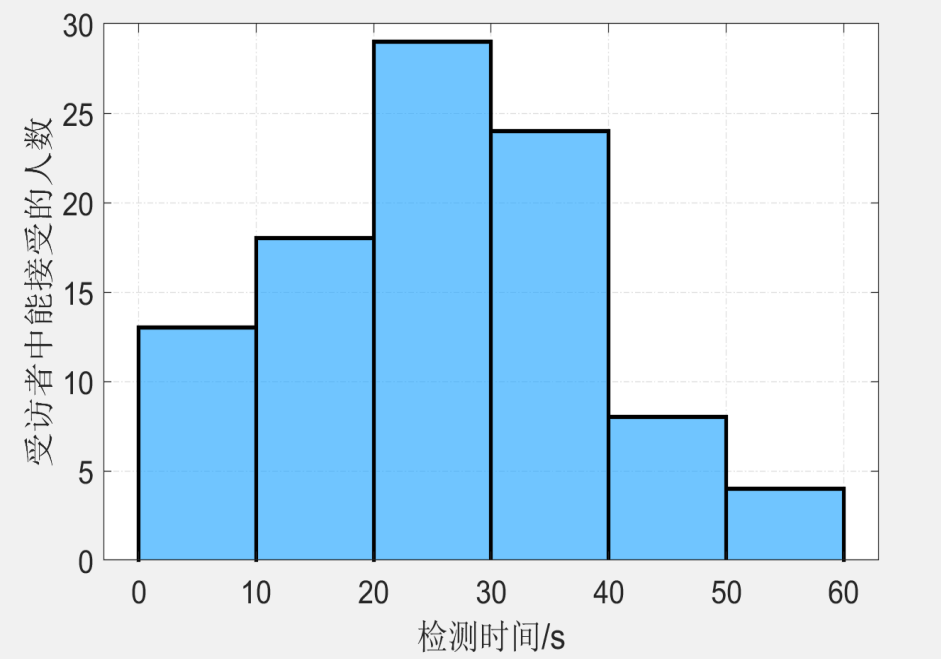

图 1.4.2.3

大部分受访者对于最大检测时间的接受程度大部分集中在 20-40s，我们团队充分考虑到这一市场需求，在选用客户端蜜罐模型时，没有考虑解混淆程度更高但耗时会大大增加的高交互性蜜罐，而是选择在低交互性蜜罐的基础上改善解混淆算法，通过算法的先进性来弥补准确性的不足，从而达到准确性与性

能的均衡。

经过我们团队的不懈努力，慧眼系统检测网站的平均时间达到了 15s 以下，与当前市面上大多数产品差别不大甚至小有优势。

同时多种检测算法的使用和最终响应式 web 页面配以各种图表的呈现形式，都使得系统更加舒适完善，更加接近市场的需求，因此本系统在一定程度上解决了当前大部分风险网站检测系统存在的覆盖面小、准确率低、误报率高的问题，有效的实现了对风险网站的精确监控。

1.5 作品特色

1.5.1 设计客户端蜜罐辅助进行恶意软件分析

与传统的服务端蜜罐不同，客户端蜜罐并不是守株待兔般的等待攻击者前来入侵。而是客户端本身去主动访问需要检测的风险网站。通过对网页上的代码进行动态执行，检测这一过程中是否有新进程产生，系统文件修改，注册表修改从而来判别网页是否含有恶意代码。

我们在对恶意软件下载类网站的检测上创新性的引入了客户端蜜罐系统来辅助检测。这一举措使得网页代码能够在一个类似沙箱的环境进行运行，同时动态执行的 JS 代码能够使得系统更加高效，准确的识别出恶意的 JS 代码，给系统带来了普通算法分析 JS 代码所无法具有的准确性。

我们发现现有的风险网站检测系统并没有使用客户端蜜罐这一技术，而我们的系统中引入这种巧妙另类的检测方法。不仅使得对恶意软件下载类的检测更加安全，同时也使得检测的准确性大大提升。

1.5.2 确立了动态解析的 JS 反混淆机制

当前许多风险网站为了逃避检测，纷纷使用 JS 混淆来逃避检测系统的检测。常见的混淆的方法有 Base64、Base95 编码、简单移位算法，甚至混淆机制还有逐渐使用复杂加密算法的趋势。如果使用单一的算法分析，必然会因算法不全面而造成恶意代码的逃逸。因此，我们确立了动态解析 JS 的反混淆机制，不管编

码得多复杂的混淆算法，经过执行，势必会触发进程件内恶意文件下载、注册表更改等敏感操作。

我们使用了 Rhino 引擎对 JS 代码进行动态执行，检测执行过程中是否存在敏感操作。通过这一思路，我们可以是的所有的 JS 混淆代码无处藏身，从而大大提高了检测的准确率同时由于 JS 代码是在类似沙箱的环境编译执行，也大大提高了系统检测的安全性。

1.5.3 基于预取的钓鱼网站检测系统

使用网路爬虫作为工具研究后发现，大型网站的拓扑结构非常复杂，网站内都有上千个和上万个链接，而钓鱼网站却出奇的简单。一般被钓鱼网站模仿的正规网站大多是银行网站，用户众多，数据量大，网站结构是进过多人团队经过长时间开发维护所形成的。钓鱼网站虽然少数页面逼真模仿正规网站，但是由于少数不法分子短时间开发部署，很难将网站拓扑复杂程度做到和正规网站相当。

针对现有的钓鱼网站检测系统主要提取单个页面特征而忽略了钓鱼网页所在网站的特征的情形，我们采用了基于网页预取的钓鱼网页检测方法，利用钓鱼网站在拓扑上的潜在弱点，结合爬虫和机器学习技术，获取并分析网站拓扑，训练得到基于网页拓扑特征的网页分类器。

1.5.4 基于自然语言处理技术的不良信息网站检测技术

传统不良信息网站检测系统是通过预先建立黑白名单来过滤不良信息网站，当用户访问不良信息网站时，根据浏览器设定的黑名单和白名单对网站进行接收或者阻挡。该名单可预先根据一定的信息建立，并且随时更新名单。随着网站数量的增长，黑白名单愈发庞大，难以管理。且不良信息网站会通过不断改变自己的 IP 地址、域名来绕过黑名单，因此传统的黑白名单技术会因地址过时而效率降低，误判率提高。

随着不良信息网站制作者反识别手段的进化，传统的过滤技术不再适用。而基于统计机器学习的过滤技术由于准确率较高、速度较快、人工成本低，成为了目前应用最广泛的技术。处于对效率和性能平衡的考虑，在多次实验后，最终决定使用了两种方案。

- TF-IDF 和 SVM 分类算法。
- 分词和朴素贝叶斯分类算法

在测试中，这两种方案有机结合下，达到了 98%的准确率和 99%的召回率。实现了对不良信息网站的高检出率、低误报率。

1.5.5 基于响应式网页设计的数据图表展示

用户在使用慧眼系统进行网站检测后后，系统将生成详细的数据图表，最终以 Web 界面的形式将报告呈现给用户和管理员。我们将使用 Chart.JS 等 JavaScript 图表库及 HTML5 等技术生成查询网站地理位置和其他数据展示。

同时，我们在系统中加入了响应式网页设计，该设计可使网站在多种浏览设备上阅读和导航，同时减少缩放、平移和滚动。因此管理人员在任何网络环境下，无论是通过桌面电脑显示器、移动电话还是其他移动产品设备都可以最舒适的方式查看并分析经过系统处理所生成的结果。

1.6 展望

本系统最大的亮点在于引入了客户端蜜罐这一针对恶意软件下载的特异性检测方法，同时针对其他不同的风险网站类型制定不同的检测方案。从而达到了较高的准确率。这是市面上第一款引入了客户端蜜罐+机器学习+黑名单三种方法进行同时识别的风险网站检测产品。通过这种全面检测的策略，使得我们的产品无论是在总体检出率和单项检出率上的表现都明显优于市面上其他同类产品。我们希望通过慧眼系统能够帮助用户及企业准确的识别出各种类型的风险网站。为我国的网络生态建设贡献出自己的一份力量。

慧眼的目标是成为全网实时工作的、强大的风险网站检测系统，我们乐于与其他检测系统开发团队、媒体和用户合作，更好的了解应用和行业的趋势。本产品适用于广大用户和企业，可以有效保障用户的健康浏览从而免受风险网站的困扰。

随着网络的日益发展以及风险网站类型的不断增多，市场对风险网站检测系

统的需求也会随之日益加剧。我们将对系统不断地进行改善优化并根据新的需求采取新技术，加入与之对应的新功能，使该系统能不断突破不断发展。

第二章 作品设计与实现

2.1 系统架构

2.1.1 系统架构分层视图

本系统分为表现层、服务层、业务逻辑层、数据访问层五层，系统总体架构图如图 2.1.1 所示：

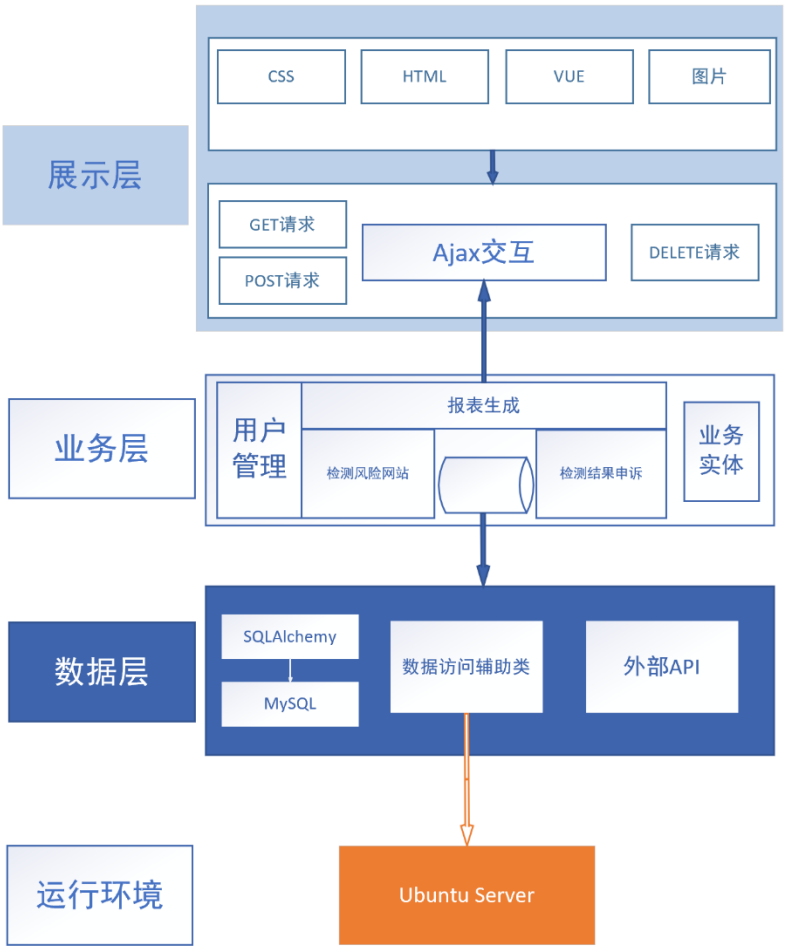

图 2.1.1 系统总体架构图

2.1.1.2 展示层

负责与用户交互，接收用户的输入，将其转化成业务层能够理解的消息格式，并调用业务层相关的业务逻辑接口来处理用户的输入，最后将得到的结果转换成某种格式呈现给用户。展示层还需要检验用户输入数据的格式，并对错误的数据输入做出响应。在 Web 应用中，展示层通常位于浏览器中。

2.1.1.3 业务层

- 业务逻辑层的责任：
  - 负责处理系统的业务逻辑
  - 负责对用户定义的流程进行建模
  - 负责数据访问层和展示层的通讯
  - 负责将错误信息返回给展示层

- 业务逻辑层的组成包括：
  - 业务实体
    - 业务实体提供对业务数据及相关功能（在某些设计中）的状态编程访问。
    - 业务实体可以是可序列化的，以保持它们的当前状态。例如，应用程序可能需要在本地磁盘、桌面数据库（如果应用程序脱机工作）或消息队列消息中存储实体数据。
    - 业务实体不直接访问数据库。全部数据库访问都是由相关联的数据访问逻辑组件提供的。
    - 业务实体不启动任何类型的事务处理。事务处理由使用业务实体的应用程序或业务过程来启动。在应用程序中表示业务实体的方法有很多（从以数据为中心的模型到更加面向对象的表示法），如通用 Data Set、自定义业务实体组件、有类型的 Data Set 和 XML。

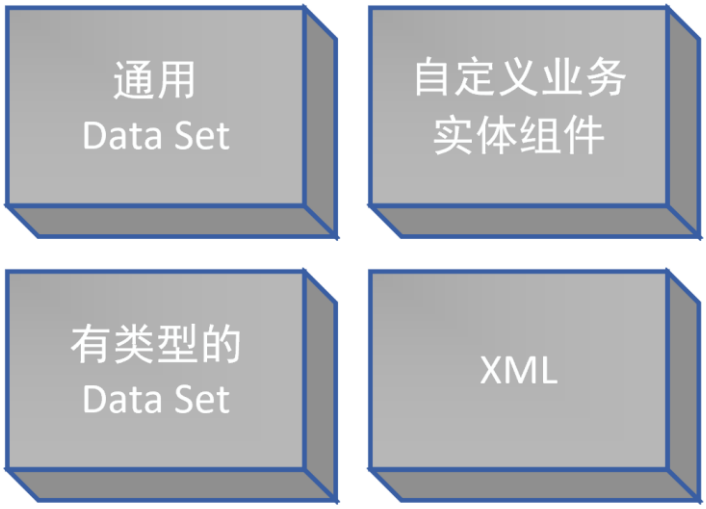

图 2.1.1.3(a)业务实体表现形式

○ 业务逻辑组件

业务逻辑组件封装业务逻辑和应用状态。业务逻辑是一种集中于实现业务规则和行为的应用逻辑，同时包括维护全局的一致性，例如数据合法性验证。业务逻辑组件应该设计成容易测试的、独立于表现层和数据访问层。本系统的业务逻辑组件可划分为如图 2.1.1.3(b)的形式。

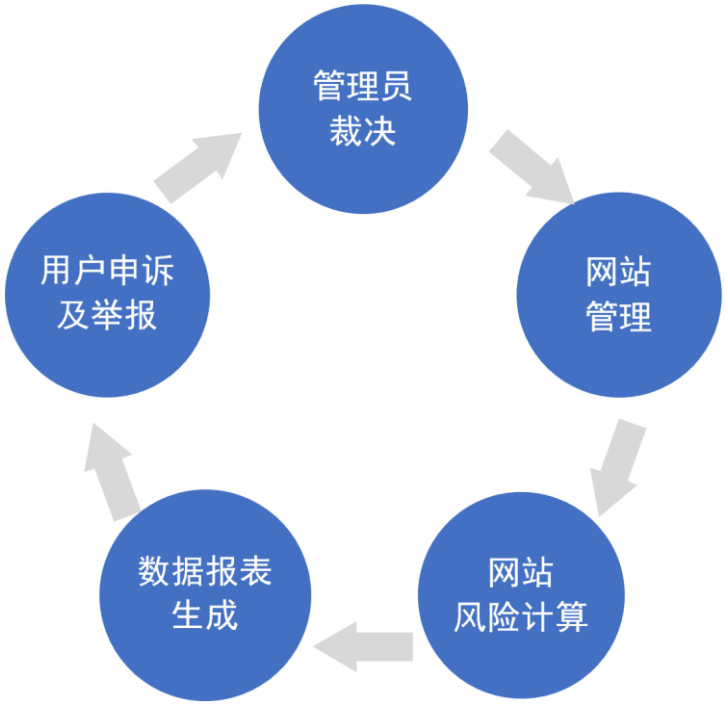

图 2.1.1.3(b) 业务逻辑组件

2.1.1.4 数据层

数据访问层：又称为 DAL 层，有时候也称为是持久层，其功能主要是负责数据库的访问。简单的说法就是实现对数据表的 Select(查询),Insert(插入), Update（更新），Delete（删除）等操作。如果要加入 ORM 的元素，那么就会包括对象和数据表之间的 mapping （映射），以及实体对象的持久化。

本系统的数据访问层使用了 Pandas 框架，将数据库表映射成数据实体，并封装其持久化的操作。出此之外，数据访问层还封装了 URLhaus Abuse，Auth0 API 的调用，通过辅助类来实现数据的本地存储

2.1.2 系统开发技术

2.1.2.1 前端开发相关技术

- HTML5

HTML5 是 HTML 下一个主要的修订版本，现在仍处于发展阶段。目标是取代 1999 年所制定的 HTML4.01 和 XHTML1.0 标准，以期能在互联网应用迅速发展的时候，使网络标准达到符合当代的网络需求。广义论及 HTML5 时，实际指的是包括 HTML、CSS 和 JavaScript 在内的一套技术组合。

本作品使用 HTML 的 Canvas API 进行二维绘图，在浏览器端进行动画渲染，实现数据可视化。

- JavaScript

JavaScript 是一种广泛用于客户端网页开发的脚本语言。它最初由网景公司的 Brendan Eich 设计，是一种动态、弱类型、基于原型的语言，内置支持类别。除前端的基本验证外，本作品还使用 JavaScript 进行动画绘制渲染，实现数据可视化。

- Vue.JS

Vue 是一套构建用户界面的渐进式框架。与其他重量级框架不同的是，Vue 采用自底向上增量开发的设计。Vue 的核心库只关注视图层，并且非常容易学习，非常容易与其它库或已有项目整合。另一方面，Vue 完全有能力驱动采用单文件

组件和 Vue 生态系统支持的库开发的复杂单页应用。比起其他重量级框架来，Vue 具有易用性，灵活性，高性能性等特点，考虑到上面种种因素，我们采用了 Vue 框架来开发我们的系统前端。

2.1.2.2 服务端开发相关技术

- Python

Python 是一种面向对象、解释型编程语言，具有近二十年的发展历史，成熟且稳定。它包含了一组完善而且容易理解的标准库，能够轻松完成很多常见的任务。它的语法简捷和清晰，尽量使用无异义的英语单词，与其它大多数程序设计语言使用大括号不一样，它使用缩进来定义语句块。

与 Scheme、Ruby、Perl、Tcl 等动态语言一样，Python 具备垃圾回收功能，能够自动管理存储器使用。它经常被当作脚本语言用于处理系统管理任务和网络程序编写，然而它也非常适合完成各种高级任务。Python 虚拟机本身几乎可以在所操作系统中运行。

2.1.2.4 自然语言处理相关技术

- TextBlob

TextBlob 是一个用 Python 编写的开源的文本处理库。它可以用来执行很多自然语言处理的任务，比如，词性标注，名词性成分提取，情感分析，文本翻译，等等。

- Jieba

Jieba 结巴中文分词是一个优秀的 Python 中文分词组件。它基于 Trie 树结构实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图(DAG)，采用了动态规划查找最大概率路径，找出基于词频的最大切分组合对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。

功能：分词、添加自定义词典、关键词提取、词性标注、并行分词。

支持三种分词模式，及繁体分词、自定义词典：

- 精确模式，试图将句子最精确地切开，适合文本分析；

- 全模式，把句子中所有的可以成词的词语都扫描出来，速度非常快；
- 搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。

- **TF-IDF 算法**

TF-IDF(term frequency-inverse document frequency) 是一种用于资讯检索与文本挖掘的常用加权技术。TF-IDF 是一种统计方法，用以评估字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。TF-IDF 加权的各种形式常被搜索引擎应用，作为文件与用户查询之间相关程度的度量或评级。

TFIDF 的主要思想是：如果某个词或短语在一篇文章中出现的频率 TF 高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。TFIDF 实际上是：TF * IDF，TF 词频(Term Frequency)，IDF 逆向文件频率(Inverse Document Frequency)。TF 表示词条在文档 d 中出现的频率。IDF 的主要思想是：如果包含词条 t 的文档越少，也就是 n 越小，IDF 越大，则说明词条 t 具有很好的类别区分能力。如果某一类文档 C 中包含词条 t 的文档数为 m，而其它类包含 t 的文档总数为 k，显然所有包含 t 的文档数 n=m+k，当 m 大的时候，n 也大，按照 IDF 公式得到的 IDF 的值会小，就说明该词条 t 类别区分能力不强。但是实际上，如果一个词条在一个类的文档中频繁出现，则说明该词条能够很好代表这个类的文本的特征，这样的词条应该给它们赋予较高的权重，并选来作为该类文本的特征词以区别与其它类文档。这就是 IDF 的不足之处. 在一份给定的文件里，词频（term frequency，TF）指的是某一个给定的词语在该文件中出现的频率。这个数字是对词数(term count)的归一化，以防止它偏向长的文件。（同一个词语在长文件里可能会比短文件有更高的词数，而不管该词语重要与否。）对于在某一特定文件里的词语来说，它的重要性可表示为：

$$tf_{i,j} = \frac{n_{i,j}}{\Sigma_k n_{k,i}}$$

以上式子中分子是该词在文件中的出现次数，而分母则是在文件中所有字词的出现次数之和。

逆向文件频率（inverse document frequency，IDF）是一个词语普遍重要性的度量。某一特定词语的 IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取以 10 为底的对数得到：

$$idf_i = lg\frac{|D|}{|\{j:t_i \in d_j\}|}$$

某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。

2.1.2.5 机器学习相关技术

- **朴素贝叶斯分类器**

朴素贝叶斯分类器是基于贝叶斯定理与特征条件独立假设的分类方法。最为广泛的两种分类模型是决策树模型(Decision Tree Model)和朴素贝叶斯模型（Naive Bayesian Model，NBM）。和决策树模型相比，朴素贝叶斯分类器(Naive Bayes Classifier 或 NBC)发源于古典数学理论，有着坚实的数学基础，以及稳定的分类效率。同时，NBC 模型所需估计的参数很少，对缺失数据不太敏感，算法也比较简单。理论上，NBC 模型与其他分类方法相比具有最小的误差率。

- **支持向量机**

支持向量机（support vector machines，SVM）是一种二分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机；SVM 还包括核技巧，这使它成为实质上的非线性分类器。SVM 的的学习策略就是间隔最大化，可形式化为一个求解凸二次规划的问题，也等价于正则化的合页损失函数的最小化问题。SVM 的的学习算法就是求解凸二次规划的最优化算法。

- **线性回归**

线归分析是一种预测性的建模技术，它研究的是因变量（目标）和自变量（预测器）之间的关系。这种技术通常用于预测分析，时间序列模型以及发现变量之间的因果关系。通常使用曲线/线来拟合数据点，目标是使曲线到数据点的距离差异最小。

线性回归是回归问题中的一种，线性回归假设目标值与特征之间线性相关，即满足一个多元一次方程。通过构建损失函数，来求解损失函数最小时的参数 w 和 b。通长我们可以表达成如下公式：

$$\hat{y} = wx + b$$

y^为预测值，自变量 x 和因变量 y 是已知的，而我们想实现的是预测新增一个 x，其对应的 y 是多少。因此，为了构建这个函数关系，目标是通过已知数据点，求解线性模型中 w 和 b 两个参数。

求解最佳参数，需要一个标准来对结果进行衡量，为此我们需要定量化一个目标函数式，使得计算机可以在求解过程中不断地优化。针对任何模型求解问题，都是最终都是可以得到一组预测值 y^ ，对比已有的真实值 y ，数据行数为 n ，可以将损失函数定义如下：

$$L = \frac{1}{n}\sum_{i=1}^{n} (\hat{y_i} - y_i)^2$$

即预测值与真实值之间的平均的平方距离，统计中一般称其为 MAE(mean square error)均方误差。把之前的函数式代入损失函数，并且将需要求解的参数 w 和 b 看做是函数 L 的自变量，可得

$$L(w,b) = \frac{1}{n}\sum_{i=1}^{n} (wx_i + b - y_i)^2$$

现在的任务是求解最小化 L 时 w 和 b 的值，即核心目标优化式为

$$(w^*, b^*) = \arg\min_{(w,b)} \sum_{i=1}^{n} (wx_i + b - y_i)^2$$

求解方式：最小二乘法(least square method)求解 w 和 b 是使损失函数最小化的过程，在统计中，称为线性回归模型的最小二乘“参数估计”(parameter estimation)。我们可以将 L(w,b) 分别对 w 和 b 求导，得到

$$\frac{\partial L}{\partial w} = 2(w\sum_{i=1}^{n} x^2 - \sum_{i=1}^{n} x_i(y_i - b))$$

$$\frac{\partial L}{\partial b} = 2(nb - \sum_{i=1}^{n}(y_i - wx_i))$$

令上述两式为 0，可得到 w 和 b 最优解的闭式(closed-form)解：

$$w = \frac{\sum_{i=1}^{n} y_i(x_i - \overline{x})}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}$$

$$b = \frac{1}{n}\sum_{i=1}^{n}(y_i - wx_i)$$

- scikit-learn

scikit-learn 是基于 Python 语言的机器学习工具。简单高效的数据挖掘和数据分析工具可供大家在各种环境中重复使用建立在 NumPy，SciPy 和 matplotlib 上开源，可商业使用 - BSD 许可证。

Scikit-learn 的基本功能主要被分为六大部分：分类，回归，聚类，数据降维，模型选择和数据预处理。

- **常用的回归**：线性、决策树、SVM、KNN ；集成回归：随机森林、Adaboost、GradientBoosting、Bagging、ExtraTrees
- **常用的分类**：线性、决策树、SVM、KNN，朴素贝叶斯；集成分类：随机森林、Adaboost、GradientBoosting、Bagging、ExtraTrees
- **常用聚类**：k 均值（K-means）、层次聚类（Hierarchical clustering）、DBSCAN
- **常用降维**：LinearDiscriminantAnalysis、PCA

2.1.2.6 病毒扫描相关技术

- Clamav

Clam AntiVirus 是一款 UNIX 下开源的 (GPL) 反病毒工具包，专为邮件网关上的电子邮件扫描而设计。该工具包提供了包含灵活且可伸缩的监控程序、命令行扫描程序以及用于自动更新数据库的高级工具在内的大量实用程序。该工具包的核心在于可用于各类场合的反病毒引擎共享库。

ClamAV 包括一个多线程扫描程序守护程序，用于按需文件扫描和自动签名更新的命令行实用程序。ClamAV 支持多种文件格式，文件和存档解包以及多种签名语言。PDF、JS、XLS、DOCX、PPT 等。

2.1.3 系统运行环境

2.1.3.1 浏览器

系统采用 HTML5、响应式网页设计等技术，在用户跨设备、跨浏览器访问时能提供当前使用环境最好的用户体验。跨平台、跨浏览器详细支持信息如下图 2.1.3.1 浏览器兼容性：

	MAC					WIN							
	CHROME	FIREFOX	OPERA	SAFARI		CHROME	FIREFOX	OPERA	IE				
	25	20	12.14	5.1	6	25	15	12	6	7	8	9	10
RGBA	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓
HSLA	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓
Box Sizing	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓
Background Size	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓
Multiple Backgrounds	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓
Border Image	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗
Border Radius	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓
Box Shadow	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓
Text Shadow	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✓
Opacity	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓
CSS Animations	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✓
CSS Columns	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✓
CSS Gradients	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✓
CSS Reflections	✓	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗
CSS Transforms	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓
CSS Transforms 3D	✓	✓	✗	✓	✓	✗	✓	✗	✗	✗	✗	✗	✓
CSS Transitions	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✓
CSS FontFace	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
FlexBox	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗
Generated Content	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓
DataURI	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓
Pointer Events	✓	✓	✗	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗
Display: table	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓

图 2.1.3.1 H5 浏览器兼容性

2.1.3.2 服务器环境

操作系统	Ubuntu 18.04 LTS
Python 环境	Python 3.609
数据库环境	MySQL 5.7.30

2.2 服务流程

慧眼系统服务流程如图 2.2 所示。

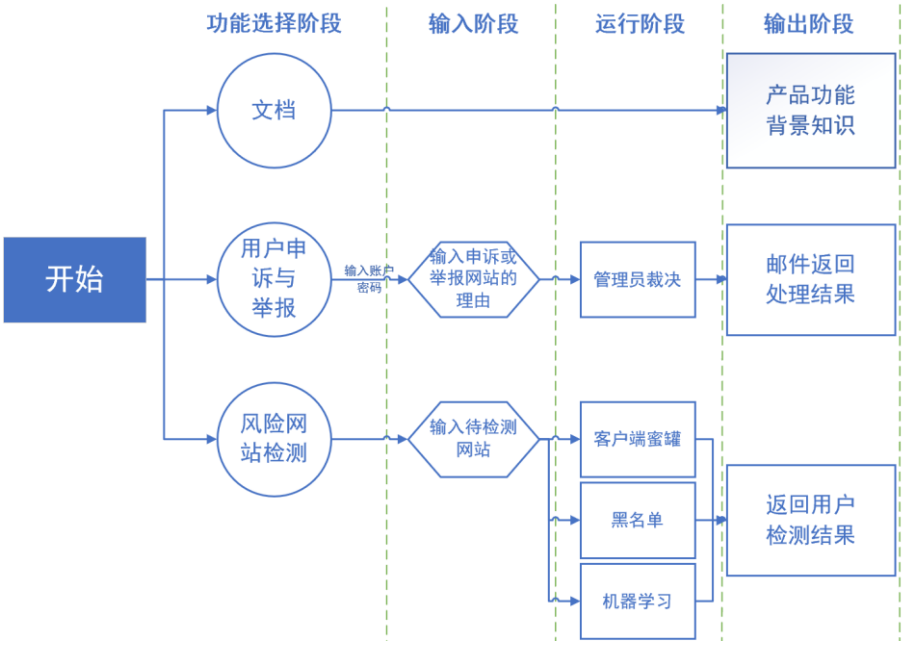

图 2.2 服务流程示意图

- **文档模块**

当进入文档界面时，会返回系统的相关信息，包括设计思路、功能、创新性等信息。

- **风险网站检测模块**

风险网站检测模块提供了三大方法(黑名单法、客户端蜜罐法、机器学习法)对用户输入的网站进行全面检测，然后返回给用户关于网站检测的相关结果和网站在地图上的具体详细信息。

- **用户申诉与举报模块**

当用户登录后，用户可输入要举报或申诉的具体网站，写清申诉原因后可进行申诉与举报。待管理员裁决后会系统会通过邮件返还用户处理结果。

2.3 功能模块

图 2.3 功能模块示意图

2.3.1 系统日志记录模块

2.3.1.1 目的

系统日志是记录系统中硬件、软件和系统问题的信息，同时还可以监视系统中发生的事件。用户可以通过它来检查错误发生的原因，或者寻找受到攻击时攻击者留下的痕迹。系统日志记录模块的作用是监控系统的运行状态，定期生成系统日志记录并以文件形式存储起来

2.3.1.2 类设计

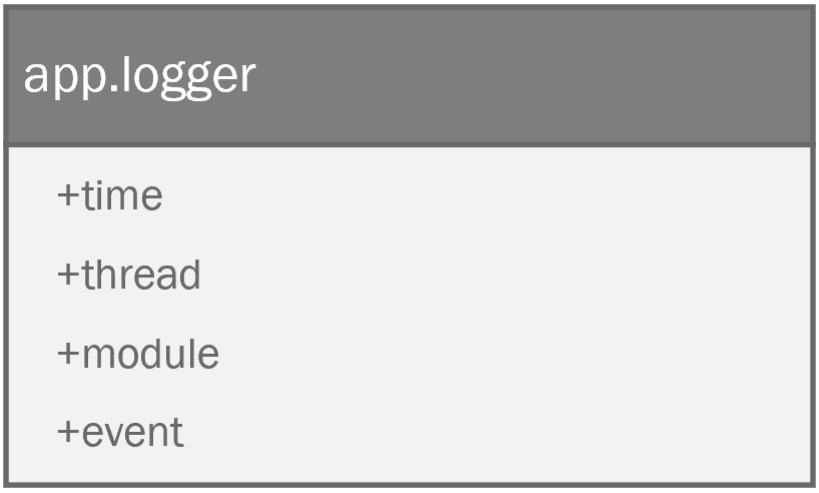

图 2.3.1.2 app.logger 类

- App.logger
  - time

    time 模块用于记录事件发生的时间
  - thread

    thread 模块用于记录发生错误的线程名称
  - module

    module 模块用于记录发生错误的模块名称
  - event

    event 模块用于记录发生错误的具体事件

2.3.2 用户管理模块

2.3.2.1 目的

实现和用户有关的基本操作，如登录，注册，删除注销用户等。

2.3.2.2 模块设计

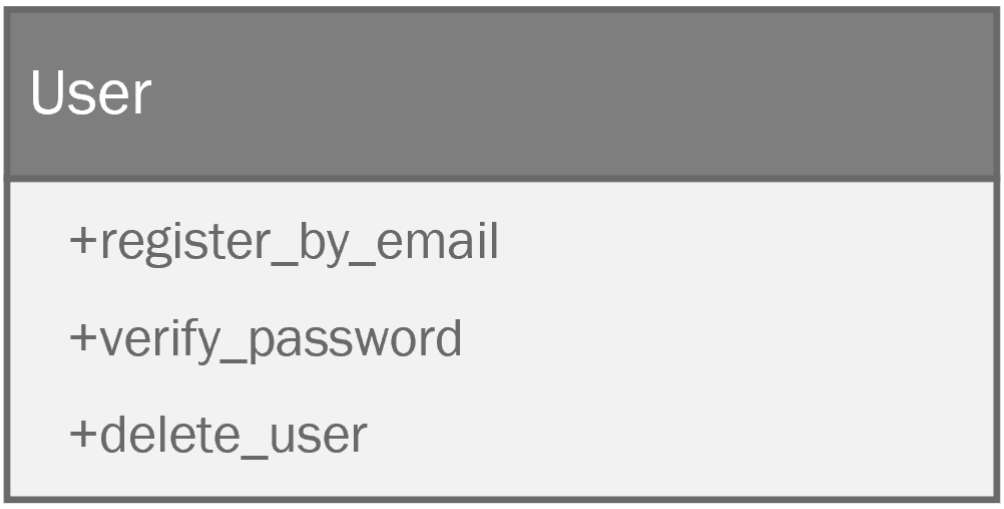

图 2.3.2.2 用户管理模块设计

- User
  - Register_by_email

    register_by_email 通过输入邮箱，用户名和密码来注册一个账号。
  - Verify_password

    verify_password 检验用户传入的账号密码是否与数据库中的数据相匹配。
  - Delete_user

    删除用户，普通用户和管理员有两种不同的调用方法，普通用户只能删除自己的账户，管理员可以通过传入 id 号来删除其他用户 3.2.2 邮件模块

2.3.3 用户申诉举报及管理员裁决模块

3.2.3.1 目的

用以处理用户的申诉举报及有关管理员裁决的相关代码逻辑 3.2.3.2 模块设计。

2.3.3.2 模块设计

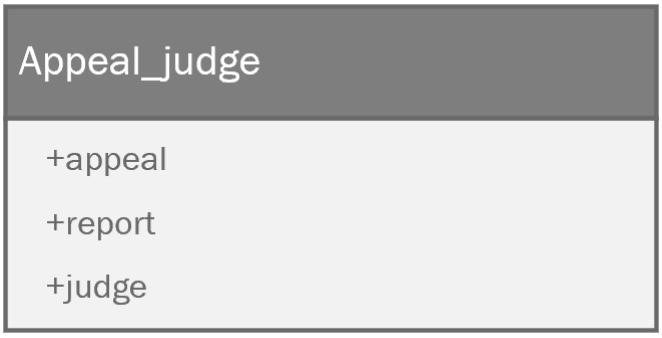

图 2.3.3.2 用户申诉举报及管理员裁决模块设计

- Appleal_judge
  - appeal

    appeal 模块用于用户对结果的申诉。
  - report

    report 模块用于用户举报某个网站。
  - judge

    judge 模块用于管理员去裁决用户举报和申诉的网站
  - Config

    config 模块从 flask 的配置信息中读取系统邮件的 SMTP 用户名，密码等信息用以发送邮件

2.3.4 邮件模块

2.3.4.1 目的

1.用户注册时给用户发送邮件使得用户可以成功注册。

2.用户申诉或举报时，可以把管理员裁决的结果通过邮件返回给用户。

2.3.4.2 模块设计

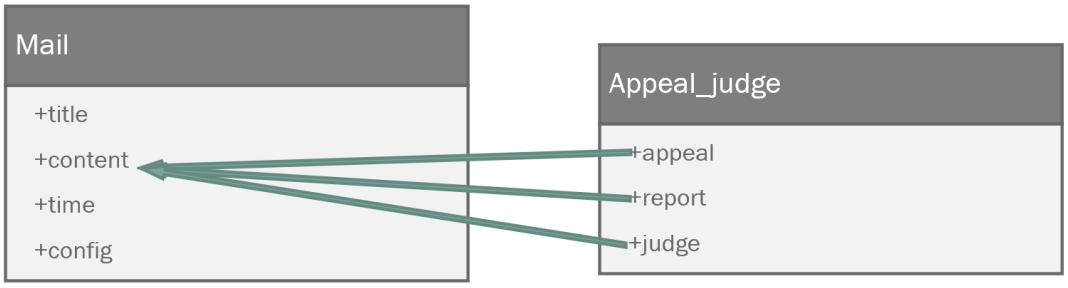

图 2.3.4.2 邮件模块设计

- Mail
  - Title

    Tittle 模块用于生成邮件的主题
  - Content

    content 从 Appeal_Judge 模块中调取用户的邮箱地址、管理员裁决结果用以生成邮件的内容
  - Time

    time 模块用以记录发送邮件的时间
  - Config

    config 模块从 flask 的配置信息中读取系统邮件的 SMTP 用户名，密码等信息用以发送邮件

2.3.5 恶意软件下载检测模块

2.3.5.1 目的

实现检测恶意软件下载网站。

2.3.5.2 实现方法

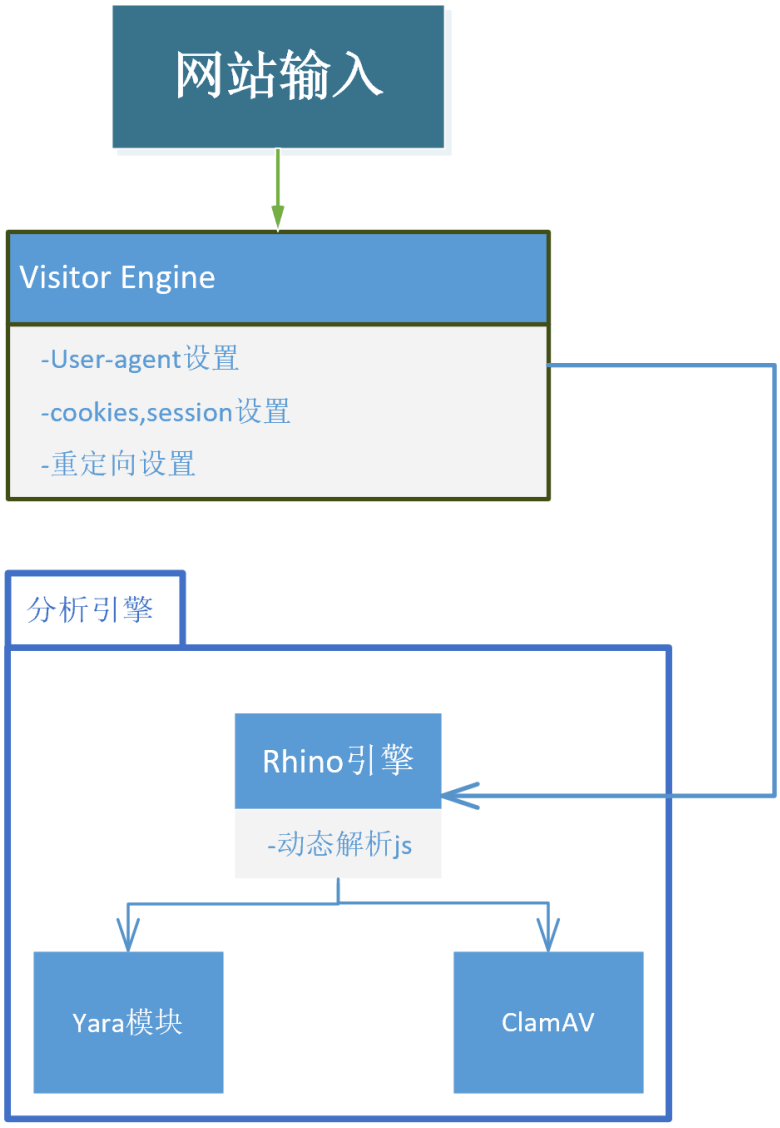

图 2.3.5.2(a) 恶意软件下载检测模块示意图

首先通过 Python 的 mechanize 模块设置 User-agent、Referer 的请求头，进行模拟浏览器访问，具体设置如下。

```
br = mechanize.Browser()
cj = cookielib.LWPCookieJar()
br.set_cookiejar(cj)
br.set_handle_equiv(True)
br.set_handle_gzip(True)
br.set_handle_redirect(True)
br.set_handle_referer(False)
br.set_handle_robots(False)
br.set_debug_responses(False)
br.set_debug_redirects(True)
br.set_handle_refresh(mechanize._http.HTTPRefreshProcessor(), max_time=0)
br.encoding = "UTF-8"
br.addheaders = [('User-Agent', honeypotconfig.useragent),('Accept', 'text/html,application/xhtml+xml,application/xml,text/javascript;
```

图 2.3.5.2(b) 设置请求头

1. 使用多种提取 JS 的方法从网页中爬取 JS，使用 Python 的 JSbeutifier 模块

JS 代码进行初步解混淆。

2. 使用 Rhino 引擎对 JS 代码进行动态执行监测过程中内存调用和文件下载情况。

3. 将内存调用情况和文件下载情况分别送给 Yara 模块和 ClamAV 模块进行病毒检测，然后返回检测结果。

2.3.6 钓鱼网站检测模块

2.3.6.1 目的

实现对钓鱼网站的检测。

2.3.6.2 实现方法

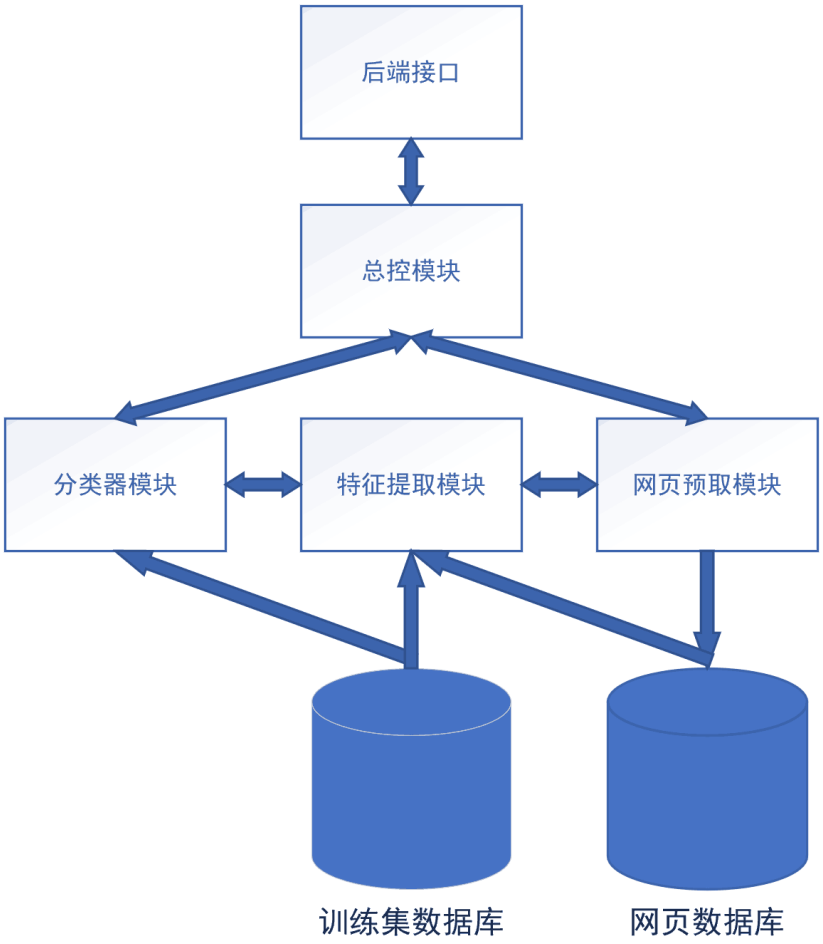

图 2.3.6.2 拓扑结构特征提取系统实现方法示意图

2.3.6.3 理论分析

我们使用爬虫作为研究工具发现大型网站的拓扑结构趋于复杂，网站内部往往有上千个网页；而一般中小型网站拓扑结构也比较复杂，网站内部也有上百个页面。但是，钓鱼网站的拓扑结果往往出奇地简单(如图 3.2.6.3)。一般被钓鱼网站模仿的正规网站用户众多、数据量大，网站结构是经过多人团队经过较长时间开发维护形成的，往往网站拓扑结构极其复杂。钓鱼网站虽然少数页面模拟得十分逼真，近似于正规网站，很难将网站拓扑复杂程度做到和正规网站相当。

基于网页预取的钓鱼网站检测方法就是利用钓鱼网站在拓扑结构上的潜在弱点，结合爬虫及机器学习技术获取并分析网站拓扑的检测流程。

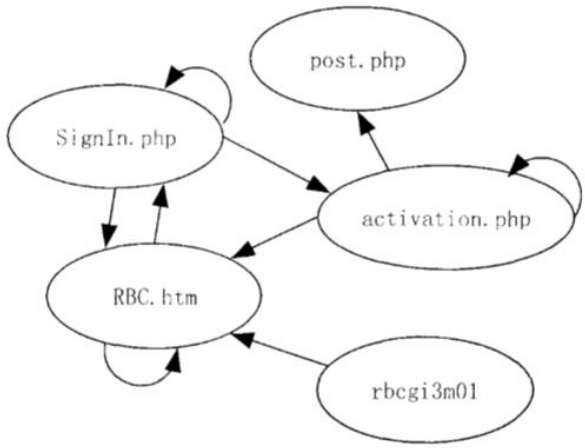

图 2.3.6.3 钓鱼网站拓扑结构趋于简单的一个例子

2.3.6.4 检测流程

1.获取初始训练集：使用爬虫采集五百个正规网站的数据抽取出特征量组成实例数据；再使用爬虫采集五百个钓鱼网站的数据抽取出特征量组成实例数据。

2.训练集的标定过程：将所有钓鱼网站实例数据中分类属性全部写为“true”，表示钓鱼网站；将所有正规网站实例苏韩剧中分类属性全部填写为“false”，表示非钓鱼网站。

3.网站拓扑特征的提取过程：包括拓扑结构构造，提取如页面平均内部链接数量、页面平均 css 文件数量、页面平均 JS 文件数量、页面平均表单数量、页面平均输入控件数量、页面平均输入密码框数量、页面表单链接数量等 15 种数值特征。

4.分类器训练过程：分类器的选择，采用增量学习方法，以及分类器参数优化。
5.对疑似钓鱼网页的检测过程：使用网络爬虫对可疑站点预取一定数量的网页，并抽取所采集的几个网页的特征数据：将抽取出的特征数据送入训练好的分类器进行分类；根据分类结果给出警告信息。

2.3.6.5 检测特点

- **高准确率**

分类问题主要的评价指标为精度和召回率，在钓鱼网站识别中，精度表示判断为钓鱼网站的所有站点中确实是钓鱼网站的比例，召回率表示所有钓鱼网站中被识别为钓鱼网站所占比例。我们采用的基于预取的钓鱼网站识别方法进行机器学习以后精度和召回率达到了 99.1%，比其他的几种钓鱼网站检测方法效果有明显提升

- **强拓展性**

由于现在抽取的特征信息种类较少(只有 15 种)，如果以后不法分子进一步提高钓鱼网站的伪装性，可以通过增加被检测网站抽取的信息种类保证准确率，同时由于采用机器学习手段进行钓鱼检测，可以不断扩充数据集，进一步提高判断的准确性。

- **较快检测速度**

传统爬虫爬取整个网页的信息速度较慢，而我们采用的基于预取的钓鱼网站检测方法改进了爬虫模块，只遍历网站部分页面，提高检测速度。

2.3.7 不良信息网站信息检测模块

2.3.7.1 目的

实现对色情、博彩等不良信息网站的检测。

2.3.7.2 算法原理说明

我们采取了两种算法去判断网是否属于不良信息网站，以及判断是何种不良

信息网站：

○ **支持向量机（support vector machines, SVM）算法**

这是一种二分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机；SVM 还包括核技巧，这使它成为实质上的非线性分类器。SVM 的的学习策略就是间隔最大化，可形式化为一个求解凸二次规划的问题，也等价于正则化的合页损失函数的最小化问题。SVM 的的学习算法就是求解凸二次规划的最优化算法。

SVM 学习的基本想法是求解能够正确划分训练数据集并且几何间隔最大的分离超平面。如下图所示，［公式］即为分离超平面，对于线性可分的数据集来说，这样的超平面有无穷多个（即感知机），但是几何间隔最大的分离超平面却是唯一的。

假设给定一个特征空间上的训练数据集

$$T = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_N, y_N)\}$$

其中，$\boldsymbol{x}_i \in \mathbb{R}^n, y_i \in \{+1, -1\}, i = 1,2, \dots N, x_i$为第$i$个特征向量，$y_i$为类标记，当它等于+1 时为正例；为-1 时为负例。再假设训练数据集是线性可分的。

几何间隔：对于给定的数据集T和超平面，定义超平面关于样本点(x_i, y_i)的几何间隔为

$$\gamma_i = y_i\left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \cdot \boldsymbol{x}_i + \frac{b}{\|\boldsymbol{w}\|}\right)$$

超平面关于所有样本点的几何间隔的最小值为

$$\gamma = \min_{i=1,2\dots,N} \gamma_i$$

实际上这个距离就是我们所谓的支持向量到超平面的距离。

根据以上定义，SVM 模型的求解最大分割超平面问题可以表示为以下约束最优化问题:

$$\max_{\boldsymbol{w},b}$$

$$\text{s.t.}\quad y_i\left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\cdot\boldsymbol{x}_i+\frac{b}{\|\boldsymbol{w}\|}\right)\geq\gamma, i=1,2,\ldots,N$$

○ **基于分词的朴素贝叶斯算法**

算法原理：朴素贝叶斯分类（NBC）是以贝叶斯定理为基础并且假设特征条件之间相互独立的方法，先通过已给定的训练集，以特征词之间独立作为前提假设，学习从输入到输出的联合概率分布，再基于学习到的模型，输入求出使得后验概率最大的输出。

我们以多分类任务为例：假设有N种可能的类别标记，即y={c1,c2,...,cN}，λij是将一个真实类别为cj的样本误分类为ci的损失，基于后验概率P(ci|cj)可获得将样本x分类为ci所产生的期望损失（expected loss），即在样本x上的“条件风险”（conditional risk）

$$R(c_i|x)=\sum_{j=1}^{N}\lambda_{i_jP(c_i|x)}$$

我们的目的是寻得一个判定准则h：X-->Y，以最小化总体风险：

$$R(h)=E_{(x)}[R(h(x)|x)]$$

对每一个样本x，若h能最小化条件风险

$$R(h(x)|x)$$

则总体风险R(h)也将被最小化，这就产生了贝叶斯判定准则（Bayes decision rule）：为最小化总体风险，只需要在每个样本上选择能使条件风险R(c|x)最小的类别标记，即

$$h^*_{(x)}=arg\max_{c\in y}R(c|x)$$

h*被称作贝叶斯最优分类器（Bayes optimal classifier），与之对应的总

体风险 R(h*)称为贝叶斯风险（Bayes risk）。1-R(h*)反映了分类器所能达到的最佳性能，即通过机器学习所能达到的模型精度的理论上限。

2.3.7.3 判断过程

- **SVM 向量机算法判断过程**

如图 2.3.7.3(a)首先利用爬虫技术爬取一定数量的色情、博彩、正常三类网站，并抽取出这三类网站能展现给用户的文字，再使用 SVM 向量机算法训练不良信息/正常网站分类器，和黄色网站/博彩网站分类器。

在需要判断网站是否属于不良信息网站，以及如果是不良信息网站，是色情网站还是博彩类网站时，就可以先后使用以上两个分类器进行判断。

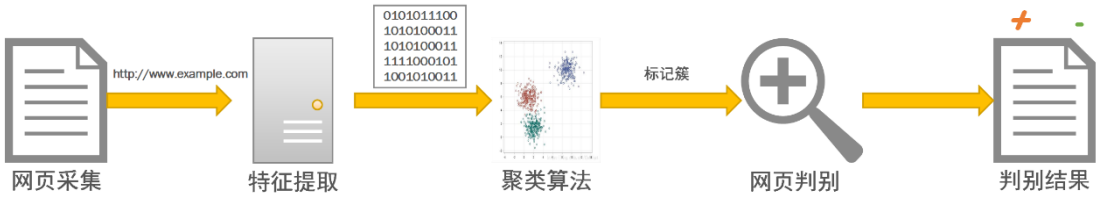

图 2.3.7.3(a) 基于 SVM 向量机算法的判断过程

- **分词的朴素贝叶斯算法判断过程**

首先利用爬虫技术爬取一定数量的色情、博彩、正常三类网站，并抽取出这三类网站能展现给用户的文字，然后使用 jieba 分词的精准分词模式将句子切割成词语列表，使用朴素贝叶斯的算法分别训练出色情/正常，博彩/正常分类器。

在需要判断网站是否属于不良信息网站，以及要确定不良信息网站类型时，就可以使用上述两个分类器的有机结合进行判断。

第三章 作品测试与分析

3.1 引言

我们设计并完成了“慧眼——基于客户端蜜罐和机器学习的风险网站检测系统”，可以针对性地对目前市面上系统地相关不足提出针对性的改进。该系统实现了黑名单法＋蜜罐检测法+机器学习三种方法多维度结合，并且为不同类型的恶意网站设计了不同的识别方案。从而大大提高了检测的准确性和针对性。下面我们针对检测的准确性、检测速度及成品的稳健性进行检测。

3.1.1 编写目的

编写测试文档的目的是通过整个测试过程，包括测试计划、测试设计、案例编写、测试总结等步骤去了解基于客户端蜜罐和机器学习的多维恶意风险网站识别系统的质量如何。是否存在缺陷，若存在缺陷则其原因是什么以及该如何修复。希望在本系统发布之前能将缺陷修复。同时向用户呈现本系统的具体测试细节，使用户能够明白系统的软件缺陷以及具体的操作方法，这将便于用户理解和使用本系统。本文档面向的读者主要是项目的开发人员和测试人员。

3.1.2 测试范围及方法

本测试对系统的各个模块先单独进行测试，再对产品整体进行测试，以保证开发人员既能将测试结果用于改进局部细节，又能对产品整体的稳健性进行完善。

- **系统功能**，我们采取模拟客户使用的动态黑盒模拟测试，从使用者体验角度出发，综合衡量系统性能和使用效果。
- **对比测试**，我们还与市面上的主流的产品进行横向对比，以直观显示出慧眼系统的优势与瓶颈，为我们在突破瓶颈时提供方向。

3.1.3 测试环境

测试环境一：云服务器

CPU	4 核
内存	8GiB
实例类型	I/O 优化
操作系统	Ubuntu 18.04 64 位
弹性网卡	eni-bp1810qgg86oxuu999g
系统盘大小	40GiB
带宽	3Mbps
Python 环境	ver3.6.9

测试环境二：PC

CPU	Intel(R) Core(TM)i5-8250U
内存	8.00GB
操作系统	Windows 10 家庭版 64 位
Web 浏览器	Chrome ver 83.0.4103.97
Python 环境	3.732
网络环境	10Mbps

3.1.4 系统可能风险

序号	终止条件	解决措施
1	系统崩溃、卡死	找出系统不稳定的原因，以及对相应模块进行修改，再重新进行测试。
2	模块集成测试中出现错误，导致有功能无法正常运行	对未正常工作的模块进行纠错、修改再重新进行集成测试以确保系统正常。
3	客户端蜜罐模块误判率过高	分析原因并修正蜜罐模块代码。
4	机器学习模块误判率过高	调整参数或适当增加特定数据集重新训练分类器。
5	前端页面展示错误	修改前端代码并重新测试。
6	服务器不稳定或出现错误	调整服务器配置或转用其他可靠服务器

3.1.5 测试结束条件

软件系统经过单元、集成、系统测试，分别达到单元、集成、系统测试的测试标准。

1. 单元测试标准

单元	结束标准
客户端蜜罐模块	正常运行，检测效率较高，检测准确率能达到 85%以上。
机器学习模块	正常运行，检测效率较高，检测准确率达 85%以上。
后端程序	流畅运行，能及时处理用户请求并运算后返回结果。
前端页面	流畅运行，提供方便且具有美感的前端交互和检测结果展示。

2. 集成测试标准

集成部位	结束标准
前端、后端之间接口	前端能将用户请求快速、准确地传递给后端处理后，后端能快速、准确地传递给前端进行展示。
后端、机器学习部分接口	机器学习能准确获得 url，在利用已训练好的分类器判断后能准确、快速返回给后端判断结果。
后端、客户端蜜罐部分接	蜜罐能准确获得 url，能利用爬虫技术

口	和引擎判断网页中各文件是否非法，然后快速返回给后端判断结果。

3. 系统测试标准

经过动态黑箱测试，使用者体验良好。系统能够结合客户端蜜罐及机器学习算法快速、准确判断 90%以上的 URL 并返回给前端进行展示。

3.2 系统功能测试过程

3.2.1 客户端功能测试

3.2.1.1 查询功能

测试编号	User-1.1
测试用例	用户正常输入正常格式 URL 进行检测
测试用例说明	验证前端能正常传递给后端用户查询的 URL，并验证能返回准确的结果
预置条件	1. 服务器处于打开状态 2. 进入检测服务主页
输入	1. 输入正常格式 URL 2. 点击确认
预期结果	正确返回判断结果
实际结果	a. 正常网页 诈骗博彩、资产敏感判断返回 false，黄色判断返回 false b. 色情网页 诈骗博彩、资产敏感判断返回 false，黄色判断返

	回 true c. 博彩诈骗等资产敏感网页 诈骗博彩、资产敏感判断返回 true，黄色判断返回 false

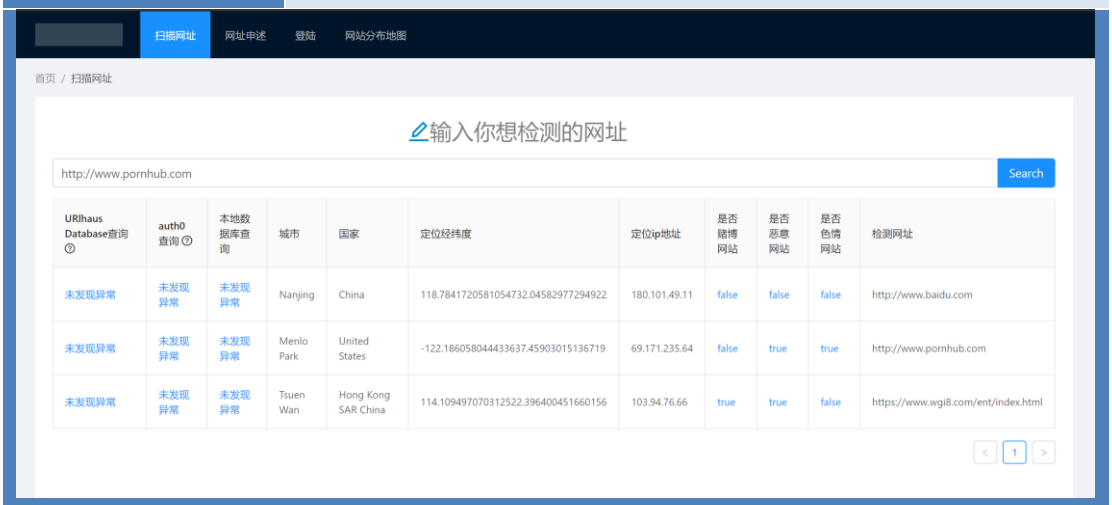

图 3.2.1.1a 查询功能正常测试

结果分析	实际结果与预期结果一致，网站查询功能运行正常

测试编号	User-1.2
测试用例	用户输入非法格式 URL 进行检测
测试用例说明	验证前、后端代码的健壮性与异常处理
预置条件	1. 服务器处于打开状态 2. 进入检测服务主页
输入	1. 输入非法格式 URL 2. 点击确认
预期结果	网站提示 URL 格式错误，且不干扰系统运行
实际结果	网站提示 URL 格式错误，且不干扰系统运行

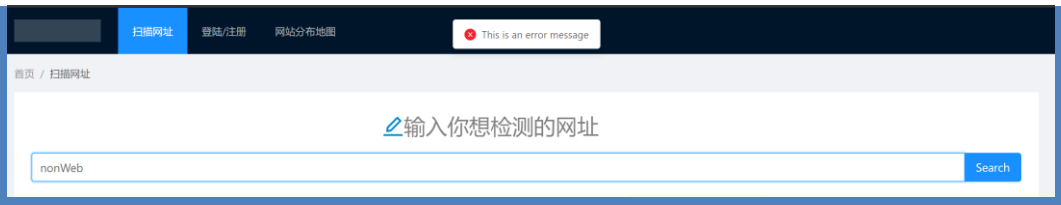

图 3.2.1.1b 查询功能非法输入测试

结果分析	实际结果与预期结果一致，在非法格式 URL 处理上的代码具有好的健壮性

3.2.1.2 用户注册

测试编号	User-2.1
测试用例	用户输入账号密码进行注册
测试用例说明	验证用户注册功能
预置条件	1.服务器处于打开状态 2.进入用户注册页面
输入	1.输入账号、密码 2.提交
预期结果	注册成功，能再次利用账号密码登录
实际结果	注册成功，能再次利用账号密码登录

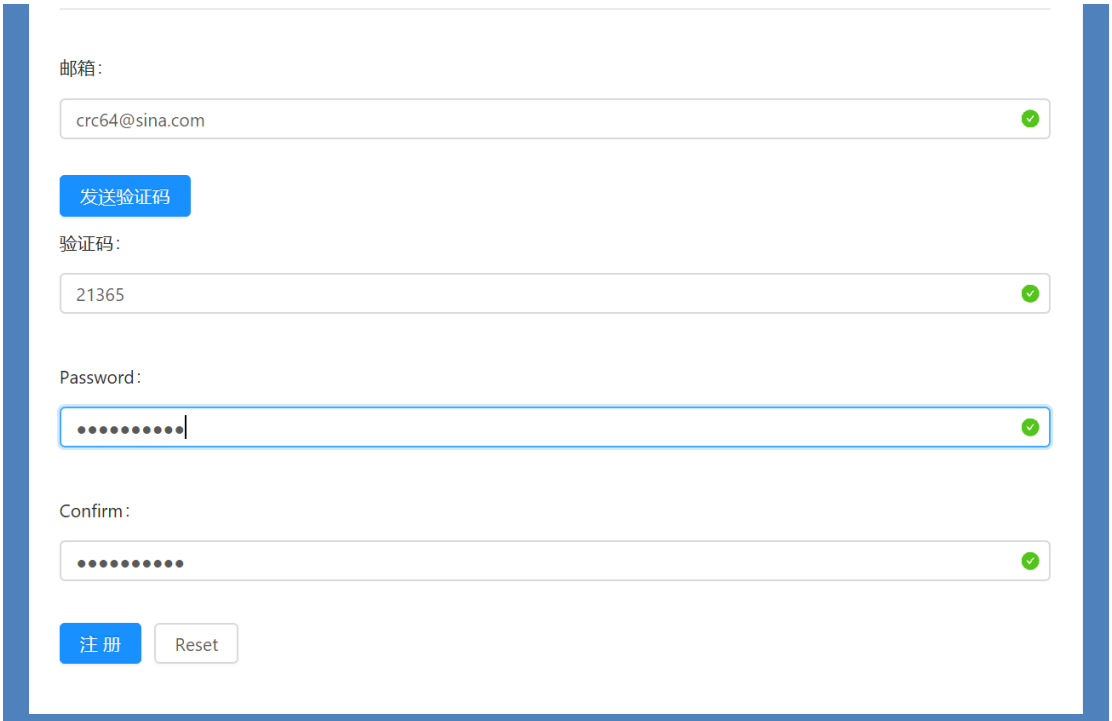

图 3.2.1.2 用户注册功能测试

结果分析	实际结果与预期结果一致，注册功能正常

3.2.1.3 网站位置分布与危险等级显示

测试编号	User-3.1
测试用例	检视网站地图页面
测试用例说明	验证网站位置查询与其前端展示的正确性
预置条件	1.服务器处于打开状态 2.进入检测服务主页查询 3.查询后进入网站地图页面检视
输入	查询几个合法与恶意网站，再点击网站地图检视页面
预期结果	显示查询过的网站位置，并且用颜色标注出其危险等级
实际结果	正确显示查询过网站的大致地理位置，并有颜色

	标注其危险等级

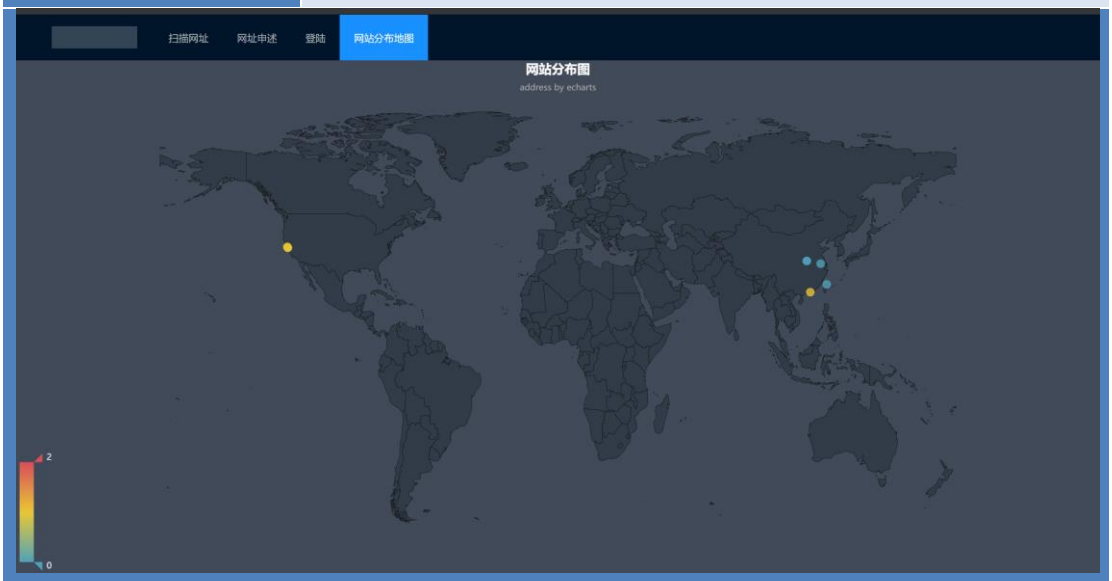

图 3.2.1.3 网站分布地图测试

结果分析	实际结果与预期结果一致，网站地图功能正常

3.2.1.4 用户申诉

测试编号	User-4.1
测试用例	申报网站分类或地理位置错误信息
测试用例说明	进入网址申报界面，填写申报信息，点击提交
预置条件	1.服务器处于打开状态 2.进入网站申诉页面
输入	填写申报信息后提交
预期结果	后台返回信息，提示已提交申诉信息
实际结果	后台返回信息，提示已提交申诉信息

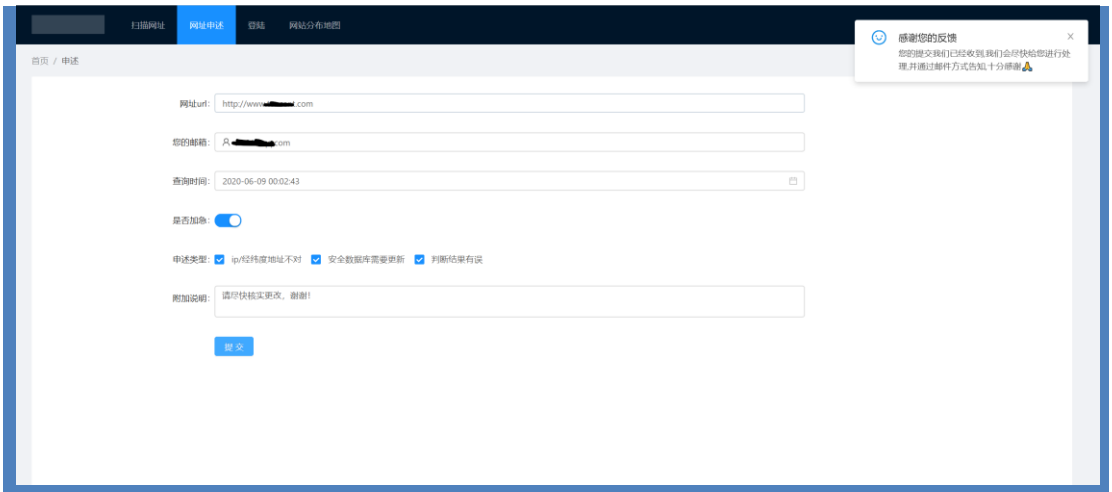

图 3.2.1.4 申诉功能测试

结果分析	实际结果与预期结果一致，申诉功能可用

3.2.1.5 管理员登录后台

测试编号	Admin-1.1
测试用例	管理员登录后台
测试用例说明	进入管理员登陆页面，输入账户和密码
预置条件	1.服务器处于打开状态 2.进入网站管理员登录页面 3.网站有存储管理员的账户
输入	按要求输入管理员账户与密码后点击提交
预期结果	导航栏出现“审批申诉”模块
实际结果	导航栏出现“审批申诉”模块

图 3.2.1.5 管理员登录后台功能测试

结果分析	实际结果与预期结果一致，管理员通过可登录进入审批申诉页面

3.2.1.6 管理员裁决

测试编号	Admin-2.1
测试用例	管理员裁决用户申诉
测试用例说明	登录管理员后进入审批申诉页面进行申诉处理
预置条件	1.服务器处于打开状态 2.进入网站管理员登录页面 3.网站有存储管理员的账户
输入	进入申诉处理页面，点击按钮进行操作
预期结果	后台成功显示出申诉内容，点击按钮可进行处理
实际结果	后台成功显示出申诉内容，点击按钮可进行处理

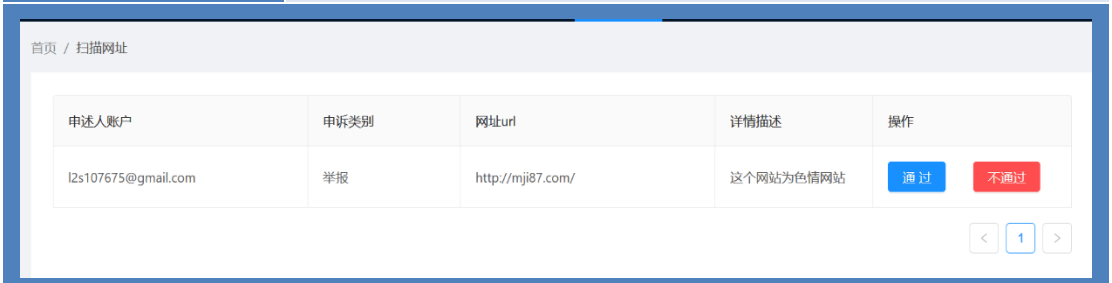

图 3.2.1.6 管理员裁决测试

结果分析	实际结果与预测结果一致，管理员裁决申诉功能正常，

3.2.1.7 产品信息安全测试

尽管我们在设计该系统时就考虑到了产品本身的信息安全，但出于客观、严谨考虑，且为了提供更稳定可靠的服务，我们还需要另外对产品本身进行一定的安全性测试，以保证系统使用者和系统的信息安全。

3.2.1.7.1 测试工具与方法

我们采用了 Acunetix Web Vulnerability Scanner 作为我们的测试工具。Acunetix 作为一款先进的 Web 漏洞扫描程序，可以独立使用，也可以作为复杂环境的一部分使用。它提供内置的漏洞评估和漏洞管理，以及许多与市场领先的软件开发工具集成的选项。

它具有以下强大的功能：

a)、自动的客本分析器,允许对 Ajax 和 web2.0 应用程序进行安全性测试

b)、业内最先进的深入的 SQL 注入和跨站脚本测试

c)、高级渗透测试工具,例如 Http Editor 和 HTTP Fuzzer

使用 Acunetix Web Vulnerability Scanner 作为测试工具，能够客观有力地检测我们产品的安全性。

3.2.1.7.2 测试结果

测试编号	Sec-1.1
测试用例	AWVS 自动化扫描漏洞
测试用例说明	使用 AWVS 测试系统是否存在安全漏洞
预置条件	1.服务器处于开启状态 2.拥有一台装有 AWVS 的个人电脑
输入	实用另一台装有 AWVS 的电脑对域名进行扫描
预期结果	扫描结果为低危
实际结果	扫描结果为低危

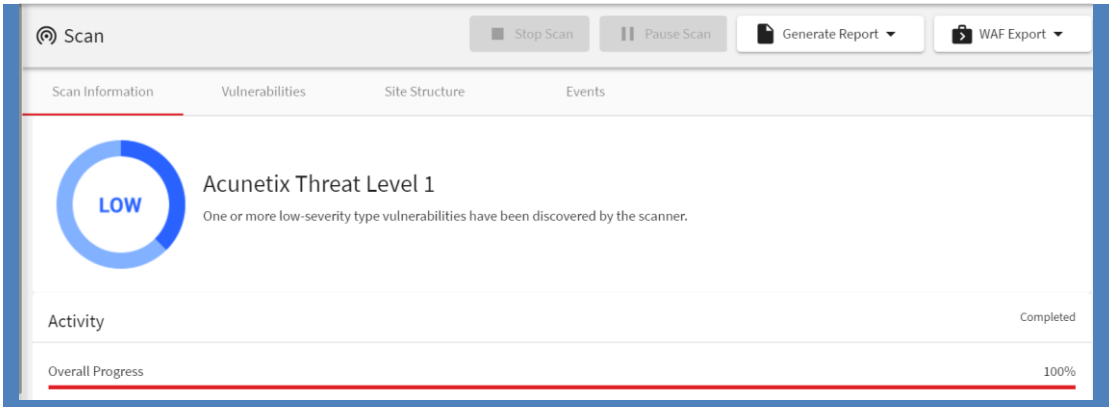

图 4.4.2 AWVS 测试结果

结果分析	实际结果与预期结果一致，系统安全性强，证明了网站本身的信息安全。

3.2.2 服务端功能测试

3.2.2.1 生成管理员账户

测试编号	Back-1.1
测试用例	生成管理员账户
测试用例说明	测试能否正常生成管理员账户
预置条件	服务器处于开启状态
输入	1. 进入服务器命令行模式 2. 执行代码文件夹中的 admin.py 文件 3. 进入 mysql 命令行 4. 执行相关数据库指令
预期结果	成功生成管理员账户
实际结果	成功生成管理员账户

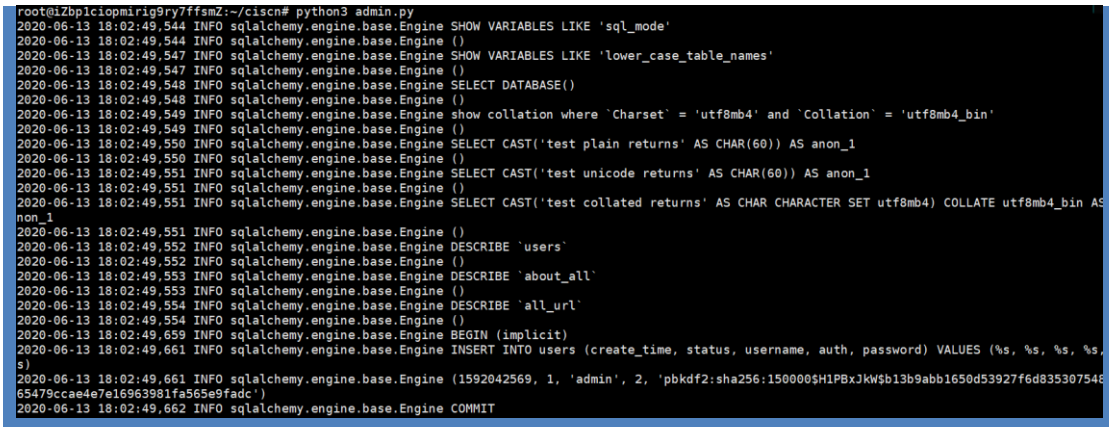

图 3.2.2.1 生成管理员账户测试

结果分析	实际结果与预期结果一致，数据库管理系统功能正常。

3.2.2.2 删除数据库中的网站

测试编号	Back-1.2
测试用例	删除数据库中的网站
测试用例说明	测试能否正常删除数据库中的网站
预置条件	服务器处于开启状态
输入	1.打开服务器端 2.进入 mysql 命令行 3.执行相关数据库指令
预期结果	成功删除该条记录

实际结果	成功删除该条记录

```
root@iZbp1ciopmirig9ry7ffsmZ:~# mysql -uroot -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 145
Server version: 5.7.30-0ubuntu0.18.04.1 (Ubuntu)

Copyright (c) 2000, 2020, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement

mysql> use ciscn;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> delete from all_url where url = 'www.baidu.com';
Query OK, 1 row affected (0.00 sec)

mysql> select * from all_url where url = 'www.baidu.com';
Empty set (0.00 sec)
```

图 3.2.2.2 删除数据库中的网站

结果分析	实际结果与预期结果一致，数据库中网站可被正常删除

3.3 对比测试

3.3.1 测试说明

我们从互联网上分别搜集了 100 个恶意软件下载网站，100 个钓鱼网址，100 个色情网站，100 个博彩网站，100 个正常网站。将这 500 个样本分别送入我们的**“基于客户端蜜罐和机器学习的风险网站识别系统”，腾讯网址安全中心，百度网址安全中心，** VirusTotals **四个不同的网站安全检测系统**中。我们将从总体检出率，假阳性率（实际为正常网站但检出为风险网站），假阴性率（实际为风险网站但检出为正常网站），分项检出率这四个不同的指标来分别衡量不同系统的优势和不足，也为接下来系统的改进提供思路。

3.3.2 总体检出率对比测试与分析

表 3.3.2 总体检出率测试

	慧眼	腾讯网址安全中心	百度网址安全中心	VirusTotals
风险网站样本数	400	400	400	400
检出数量	357	268	279	129
未检出数量	43	132	121	271
检出率	89.25%	67%	69.75%	32.25%

从表 5.2.1 可以看出，我们的系统在检出率上远超其余的三个系统，对其他几个系统均有显著性的优势。腾讯网址安全中心和百度网址安全中心检出率相近，均为 70%左右，这主要是由于他们两家拥有广大的用户，可以根据广大的用户反馈来添加黑名单的结果。在四种不同的系统中，VirusTotals 的检出率最低，这主要是因为 VirusTotals 主要针对的是恶意软件下载这种类型的风险网站进行检测，对其他类型的风险网站几乎没有检测能力。

3.2.3 恶意软件下载网站检出对比测试与分析

表 3.2.3 恶意软件检出率测试

	慧眼	腾讯网址安全中心	百度网址安全中心	VirusTotals
风险网站样本数	100	100	100	100
检出数量	80	24	28	89

未检出数量	20	76	72	11
检出率	80%	24%	28%	89%

在恶意软件检测方面，可以看到 VirusTotals 的检出率最高，主要原因是 VirusTotals 集成了 60 多种不同的引擎，在恶意软件检测方面较为擅长。而腾讯网址安全中心和百度网址安全中心在这一方面表现较差，检出率均仅为 30%不到，主要原因是大多数恶意软件下载网站生命周期较短，仅为几个小时不到，此时腾讯网址安全中心和百度网址安全中心所广泛采用的黑名单法则表现较差。

而我们使用的客户端蜜罐+机器学习检测恶意软件下载网站，虽然检出率达不到 VirusTotals 的 89%，但也达到了较为优秀的 80%。如果要进一步提升检出率，可以考虑在以后添加除了 ClamAV 的病毒扫描引擎，运用多引擎解析来提高检出率。

3.2.4 钓鱼网站检出对比测试与分析

表 3.2.4 钓鱼网站检测率测试

	慧眼	腾讯网址安全中心	百度网址安全中心	VirusTotals
风险网站样本数	100	100	100	100
检出数量	91	85	77	11
未检出数量	9	15	23	89
检出率	91%	85%	77%	11%

从表 5.2.2.2，可以看到我们为钓鱼网站所专门使用的基于预取的钓鱼网站检测法取得了优秀的效果，检出率达到了 91%。而腾讯网址安全中心在这方面表

现得也不错，值得注意的是，在 100 个钓鱼网站样本中，有 30 个是对腾讯产品的伪冒，而腾讯网址安全中心对这 30 个样本实现了 100%检出，由此可以看出，腾讯对自家产品的相关检测能力较为优秀。而 VirusTotals 由于只能针对恶意软件下载网站进行分析，在这里表现不佳，检测出的 11 个样本为恰好拥有恶意软件的样本。

3.2.5 色情网站检出对比测试与分析

表 3.2.5 色情网站检测率测试

	慧眼	腾讯网址安全中心	百度网址安全中心	VirusTotals
风险网站样本数	100	100	100	100
检出数量	97	73	85	20
未检出数量	3	27	15	80
检出率	97%	73%	85%	20%

色情网站检测是慧眼系统表现得最好的一个方面，检出率达到了 97%，我们通过爬取外网色情平台上的关键词利用文本聚类等相关技术获取相关的关键词获取色情关键词列表，再去除一些生活中的常用分词来使得分词算法更加真实，从而建立起一个最真实的语料库。由于使用了这一技术，使得我们系统对色情网站有着极高的检出率。

VirusTotals 系统对色情网站的检测是其在非恶意软件下载类网页中表现最好的，原因可能是色情网站很大概率也是携带各种恶意跳转，恶意下载。

3.2.6 博彩赌博网站检出对比测试与分析

表 3.2.6 博彩赌博类网站检测率测试

	慧眼	腾讯网址安全中心	百度网址安全中心	VirusTotals
风险网站样本数	100	100	100	100
检出数量	89	86	89	9
未检出数量	11	14	11	91
检出率	89%	86%	89%	9%

从表 5.2.2.4 可以看出，对于博彩网站，慧眼系统，腾讯网址安全中心和百度网址安全中心表现均较为优秀。这主要是因为博彩网站相当大一部分都带有色情信息，而容易同时被色情网站的识别算法所识别出来。但比起腾讯网址安全中心和百度网址安全中心，慧眼系统拥有独特的博彩网站识别算法。算法中主要加入了对正常网站的数据测验集。由于金融系统的特殊性，很多小众但合法的金融网站容易被误判为风险网站。这一点在腾讯网址安全中心和百度网址安全中心上均没有被妥善解决，而我们的系统可以根据正常的数据集不断学习适当的修改模型，这在很大程度上减少了误报。

3.2.7 假阳性率对比测试与分析

表 3.2.7 假阳性率测试

	慧眼	腾讯网址安全中心	百度网址安全中心	VirusTotals

正常网站样本数	100	100	100	100
正确判断次数	92	82	86	98
误判次数	5	18	14	2
假阳性率	5%	18%	14%	2%

从表 5.2.3 可以看出，假阳性方面，表现得最优秀的是 VirusTotals 系统，这是由于其特异性针对恶意软件下载网页，覆盖面较小的原因。在其他三种检测系统中，慧眼系统的表现也要明显优于其他两款系统。造成此种区别的主要就是在博彩赌博的检测上，腾讯网址安全中心和百度网址安全中心由于缺乏针对性的算法且一味追求效率，造成对网页检测的误报率较高。这在生活中也屡见不鲜，互联网上很多个人站长就一直在反映被腾讯或百度误报的问题，而我们的系统通过设计针对性的算法，有效的解决了这一问题，把误报率降低到 5%，适合于大规模使用。

3.4 测试总结与改进空间

3.4.1 总结

经过我们数次缜密的测试，我们将 bug 出现的概率降到最低，并通过测试验证了我们预想的设计已完全实现：

- 前后端分离，可插拔性高，bug 概率被降到最低
- UI 界面宜人、友好，运行流畅，结果展示直观
- 结合客户端蜜罐、机器学习、黑名单等方法有机结合综合判断网站安全性的效果显著。检测综合全面，不留短板；检测速度快，方便实用
- 用户登录、后台管理、申诉以及申诉审批过程、邮件回复审批结果等十分流畅，体验良好

- 检测、防止非正常输入的机制起效，系统在容错性和代码健壮性良好
- 系统本身有较高信息安全的水准
- 与同类产品相比，具有总体检出率、恶意软件下载网站检出率、钓鱼网站检出率、色情网站检出率、博彩赌博网站检出率高等显著优势

3.4.2 改进空间

- 表现形式上还可以更加丰富，由于前后端分离、可插拔的优良特性，之后我们还可以方便、快捷地进行完善
- 该系统的部分功能还可以提供 API 给其他网站或平台（如不良信息识别模块可以用在论坛网站对不良信息出现进行警告），从而进一步增加我们慧眼系统的适用性和使用范围
- 我们选择的算法模型还可以根据具体领域应用需求调整参数和数据集，来达到动态适应特定场景的效果

第四章 创新性说明

4.1 系统性创新

在我们的整个慧眼系统的搭建中，我们始终瞄准着目前行业痛点——风险网站类型和技术日益趋于纷繁复杂，识别难度大，现有安全体制过多依赖人工举报和维护黑名单。

图 4.1 全面合作式检测

网络安全事件中，核心问题往往是“**木桶效应**”。我们清楚地意识到，想要识别并精准打击恶意网站，必须尽可能多方面多角度地覆盖恶意网站类型，所以必须引入具体入微的、稳健强效的、相互合作的**多模块检测系统**。除此之外，我们还**结合了传统黑名单、人工举报**的方式，以及**宜人的交互设计**来提供服务。

黑产者布置了一个网站，或许能靠着刻意构造的复杂拓扑结构绕开预取网页检测，但很难抵挡 SVM 向量机、朴素贝叶斯对于违法文字信息的检测。即使黑产者采取了钓鱼网站式的伪装，当它想通过恶意下载或其他方式来危害用户时，我们精心构造的客户端蜜罐以及多种 JS 反混淆机制也能够有力粉碎恶意企图！

不仅如此，慧眼系统在构造时就瞄准了还需具有**高的可靠性、高效性、易用性、可移植性**的优点。在后续实用中，可灵活调整慧眼系统模块的配置，**达到高度定制化的实用效果**。

4.2 模块内创新

4.2.1 设计客户端蜜罐辅助进行恶意软件分析

与传统的服务端蜜罐不同，客户端蜜罐并**不是守株待兔**般的等待攻击者前来入侵。而是客户端本身去**主动访问**需要检测的风险网站。通过对网页上的代码进行动态执行，检测这一过程中是否有新进程产生，系统文件修改，注册表修改从而来判别网页是否含有恶意代码。

我们在对恶意软件下载类网站的检测上创新性的引入了客户端蜜罐系统来辅助检测。这一举措使得网页代码能够在一个类似沙箱的环境进行运行，同时动态执行的 JS 代码能够使得系统更加高效，准确的识别出恶意的 JS 代码，给系统带来了普通算法分析 JS 代码所无法具有的准确性。

我们发现现有的风险网站检测系统并没有使用客户端蜜罐这一技术，而我们的系统中引入这种巧妙另类的检测方法。不仅使得对恶意软件下载类的检测更加**安全**，同时也使得检测的**准确**性大大提升。

4.2.2 确立了动态解析的 JS 反混淆机制

当前许多风险网站为了逃避检测,纷纷使用 JS 混淆来逃避检测系统的检测。常见的混淆的方法有 Base64、Base95 编码、简单移位算法，甚至混淆机制还有逐渐使用复杂加密算法的趋势。如果使用单一的算法分析，必然会因算法不全面而造成恶意代码的逃逸。因此，我们确立了**动态解析 JS 的反混淆机制**，不管编码得多复杂的混淆算法，经过执行，势必会触发进程件内恶意文件下载、注册表更改等敏感操作。同时由于 JS 代码是在类似沙箱的环境编译执行，也大大提高了**系统检测的安全性**。

4.2.3 基于预取的钓鱼网站检测系统

采用基于网页预取的钓鱼网页检测方法，利用钓鱼网站在拓扑上的潜在弱点，结合爬虫和机器学习技术，获取并分析网站拓扑，**训练得到基于网页拓扑特征的网页分类器**。使得绝大多数一般黑产钓鱼网站均逃不过它的检测，配合其它检测方案，检测效果、准确性更加优异。

4.2.4 基于自然语言处理技术的不良信息网站检测技术

传统不良信息网站检测系统是通过预先建立黑白名单来过滤不良信息网站，当用户访问不良信息网站时，根据浏览器设定的黑名单和白名单对网站进行接收或者阻挡。该名单可预先根据一定的信息建立，并且随时更新名单。随着网站数量的增长，黑白名单愈发庞大，难以管理。且不良信息网站会通过不断改变自己的 IP 地址、域名来绕过黑名单，因此传统的黑白名单技术会因地址过时而效率降低，误判率提高。在 **TF-IDF+SVM 分类算法，分词+朴素贝叶斯分类算法，以及采取增量式学习模式**的“组合拳”下，我们能**动态地适应**当下不良信息网站的特征信息，使得我们的检测水准能**自动地与时俱进**而非逐渐变得不适应。

4.2.5 基于响应式网页设计的数据图表展示

用户在使用慧眼系统进行网站检测后后，系统将生成详细的数据图表，最终以 Web 界面的形式将报告呈现给用户和管理员。我们将使用 Chart.JS 等 JavaScript 图表库及 HTML5 等技术生成查询网站地理位置和其他数据展示。

同时，我们在系统中加入了**响应式网页设计**，该设计可使网站在多种浏览设备上阅读和导航，同时减少缩放、平移和滚动。因此管理人员在**任何网络环境下**，无论是通过桌面电脑显示器、移动电话还是其他移动产品设备都可以最舒适的方式查看并分析经过系统处理所生成的结果。

第五章 总结

5.1 系统设计与开发

“慧眼——基于客户端蜜罐和机器学习的风险网站检测系统”时我们Intelligence团队历时半年，以风险网站日益猖獗为背景，针对当前风险网站检测系统存在的诸多不足（如误报率高、检测精度低、解混淆能力不够），而开发的具有实用性、高效性、可靠性、灵活性的风险网站检测系统。

设计本系统时，团队充分考虑学习了风险网站泛滥的大环境以及现行的风险网站检测系统后，对现有的风险网站检测系统经过搜集整理并对其所用技术进行研究分析。从中我们找出了这些系统各自的优势以及它们存在的需要改进的问题。针对这些不足点，我们分别采用不同技术对其逐一解决优化。同时我们在自己的系统中加入与以往不同的新的检测观念以达到更为合理准确的检测效果。本系统的开发在原有风险网站检测系统基础上主要涉及以下几方面新的知识点，分别解决市场上现有系统存在的各种不足：

1. 设立客户端蜜罐辅助进行恶意软件分析
2. 确立了动态解析的JS反混淆机制
3. 基于预取的钓鱼网站检测系统
4. 基于自然语言处理技术的不良信息网站检测技术
5. 基于响应式网页设计的数据图表展示

虽然本产品已经是能够投入使用的优秀产品，但是后续仍然有很大的改善空间，我们打算对产品的算法及设计思想进行持续性的改善及调整，争取进一步提高检测的准确率。同时，我们的识别算法并不仅仅只能对单网站进行分析，在后续的改进中，我们将尝试将其接入微博、直播网点等公共平台，争取实现对于平台类网点的持续性检测。

随着网络的日益发展以及风险网站类型的不断增多，市场对风险网站检测系统的需求也会随之日益加剧。我们将对系统不断地进行改善优化并根据新的

需求采取新技术，加入与之对应的新功能，使该系统能不断突破不断发展。

参考文献

[1]CNCERT 互联网安全威胁报道 2020 年 3 月期

[2]沙泓州,刘庆云,柳厅文等恶意网页识别研究综述[J].计算机学报,2016,39(3):529-542.

[3]中国互联网信息中心.第 44 次中国互联网络发展状况统计报告[R].北京:CNNIC,2019.

[4]诸葛建伟,唐勇,韩心慧,段海新.蜜罐技术研究与应用进展.

[5]王正琦,冯晓兵,张驰.基于两层分类器的恶意网页快速检测系统研究[J].网络与信息安全学报,2017(8):44-60.

[6]基于主成分分析和随机森林的恶意网站评估与识别[J].陈远,王超群,胡忠义,吴江.数据分析与知识发现. 2018(04)

[7]基于异常特征的钓鱼网站 URL 检测技术[J].黄华军,钱亮,王耀钧.信息网络安全.2012(01)

[8]基于聚类改进的 KNN 文本分类算法[J].周庆平,谭长庚,王宏君,湛淼湘.计算机应用研究.2016(11)

[9]基于 URL 特征的 Phishing 检测方法(英文)[J].曹玖新,董丹,毛波,王田峰. Journal of Southeast University(English Edition).2013(02)

[10]基于机器学习的网页恶意代码检测方法[J].李洋,刘飚,封化民.北京电子科技学院学报.2012(04)

[11]融合域名注册信息的恶意网站检测方法研究[J].陈庄,刘龙飞. 计算机光盘软件与应用.2015(01)

[12]基于嵌套 EMD 的钓鱼网页检测算法[J].曹玖新,毛波,罗军舟,刘波.计算机学报. 2009(05)

[13]基于学习的恶意网页智能检测系统[D].王松.南京理工大学 2011

[14]安全使者 百度网址安全中心解密[J].宋辉.计算机与网络.2017(13)

[15]网络钓鱼变种解析[J].焦仃.计算机与网络.2017(24)

[16]一种基于机器学习的网页分类技术[J].孙靖超.信息网络安全.2017(09)

[17]基于蜜罐网络的邮件捕获系统分析与部署[J].李秋锐.信息网络安全.2012(01)

[18]Multiple classification of the force and acceleration signals extracted during multi ple machine processes: part 2 intelligent control simulation perspective[J]. James M. Griffin,Alejandro J. Doberti,Valbort Hernández,Nicolás A. Miranda,Maximiliano A. Vélez. The International Journal of Advanced Manufacturing Technology. 2017(9-12)

[19] M.Nawrocki,M.Wahlisch,T.C.Schmidt,C.Keil,and W.Matthias, “A Survey on Ho neypot Software and Data Analysis,”arXiv preprint arXiv:1608.06249, 2016.

[20]A.Mairh,D. Barik, K. Verma, and D. Jena, “Honeypot in network security: a surv ey,” in Proceedings of the 2011 International

[21]Conference on Communication, Computing & Security. ACM,
2011, pp. 600-605.

[22] Angelo Del’Aera, “Welcome to Thug’s documentation! – Thug 0.8.33 documenta tion,” 2017. [Online]. Available: https://buffer.github.io/thug/doc/index.html. [Access ed: 10-Apr2017]

[23]Alexandre norman, “pyClamd: Clamav with python,” 2016. [Online]. Available: h ttp://xael.org/pages/pyclamd-en.html. [Accessed: 9May2017].

[24]“Chapter 1. First steps,” 2016. [Online]. Available:
https://www.virtualbox.org/manual/ch01.html. [Accessed: 9May2017].

[25] Yin W, Kann K, Yu M, et al. Comparative Study of CNN and RNN for Natural L anguage Processing[J]. 2017.

[26]Wen Y, Zhang W, Luo R, et al. Learning text representation using recurrent convolutional neural network with highway layers[J]. 2016.

[27]Johnson R, Zhang T. Effective Use of Word Order for Text Categorization with Convolutional Neural Networks[J]. Eprint Arxiv, 2014.

[28]Canfora G, Medvet E, Mercaldo F, et al. Detection of Malicious Web Pages Using System Calls Sequences[M]// Availability, Reliability, and Security in Information Systems. Springer International Publishing, 2014.

[29]Sahoo D, Liu C, Hoi S C H. Malicious URL Detection using Machine Learning: A Survey[J]. 2017.

[30]Nicomette V, Kaâniche M, Alata E, et al. Set-up and deployment of a high-interaction honeypot: experiment and lessons learned. Journal in Computer Virology, 2011, 7(2): 143-157.

[31]Blum, Aaron, Brad Wardman, Thamar Solorio, and Gary Warner. Lexical feature based phishing URL detection using online learning//Proceedings of the 3rd ACM Workshop on Artificial Intelligence and Security(AISEC). Chicago, USA, 2010:54-60. [

[32]Chou, Neil, Robert Ledesma, Yuka Teraguchi, and John C. Mitchell. Client-Side Defense Against Web-
Based Identity Theft//Proceedings of the 11th Annual Network & Distributed System Security Symposium (NDSS). San Diego,USA. 2004:1-16.

[33]Zhang, Junjie, Christian Seifert, Jack W. Stokes, and Wenke Lee. Arrow: Generating signatures to detect drive-
by downloads// Proceedings of the 20th international conference on world wide web (WWW). Hyderabad, India, 2011:187-196.

[34]Eshete, Birhanu, Adolfo Villafiorita, and Komminist
Weldemariam. Binspect: Holistic analysis and detection of
malicious web pages. Lecture Notes of the Institute for Computer
Sciences Social Informatics & Telecommunications Engineering,
2013, 106:149-166.

[35]Canali, Davide, et al. Prophiler: a fast filter for the large-
scale detection of malicious web pages//Proceedings of the 20th international confere

nce on world wide web (WWW). Hyderabad, India, 2011: 197-206. [
[36]Liang, Bin, Jianjun Huang, Fang Liu, Dawei Wang, Daxiang Dong, and Zhaohui Liang. Malicious Web Pages Detection Based on Abnormal Visibility Recognition//Proceedings of the International Conference on E-
Business and Information System Security(EBISS'09). Wuhan, China, 2009:1-5. [
[37]Kals S, Kirda E, Kruegel C, et al. Secubat: a web vulnerability scanner//Proceedings of the 15th international conference on World Wide Web(WWW). New York, USA, 2006: 247-256.
[38]Mahmoud K, Youssef I, Andrew J. Phishing Detection: A Literature Survey. IEEE Communication Surveys & Tutorials, 2013, 15(4): 2091-2121.
[39]Paul K, Georgia K, Hector G. M. Fighting Spam on Social Web Sites A Survey of Approaches and Future Challenges. IEEE Internet Computing, 2007, 11(6): 36-45. [
[40]Kolbitsch C, Livshits B, Zorn B, et al. Rozzle: De-cloa 慧眼
internet malware//Proceedings of the IEEE Symposium on
Security and Privacy (SP). San Francisco, USA, 2012: 443-457.
[41]Thomas, Kurt, et al. Design and evaluation of a real-time url spam filtering service//Proceedings of the IEEE Symposium on Security and Privacy (SP). Oakland, California, 2011:447-462.