2020 年全国大学生信息安全竞赛

作品简介

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 日

作品简介

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。

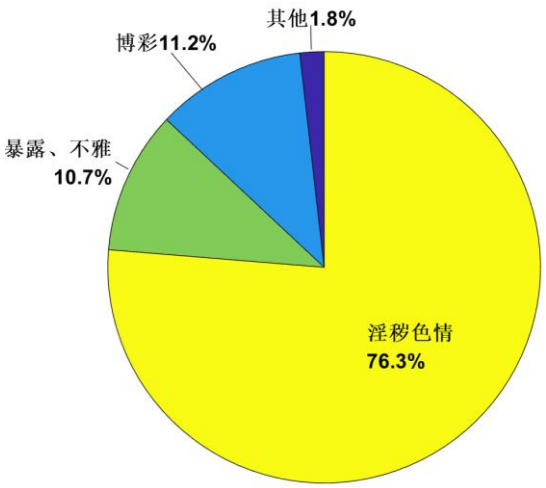

图 1.1.1(a) 2019 年各类不良信息网站占比统计

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。

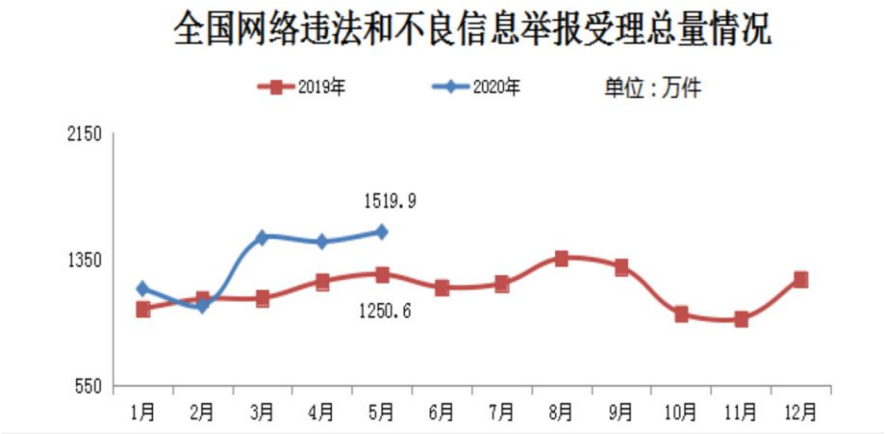

图 1.1.1(b) 2019 和 2010 年全国违法和不良信息举报受理情况

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

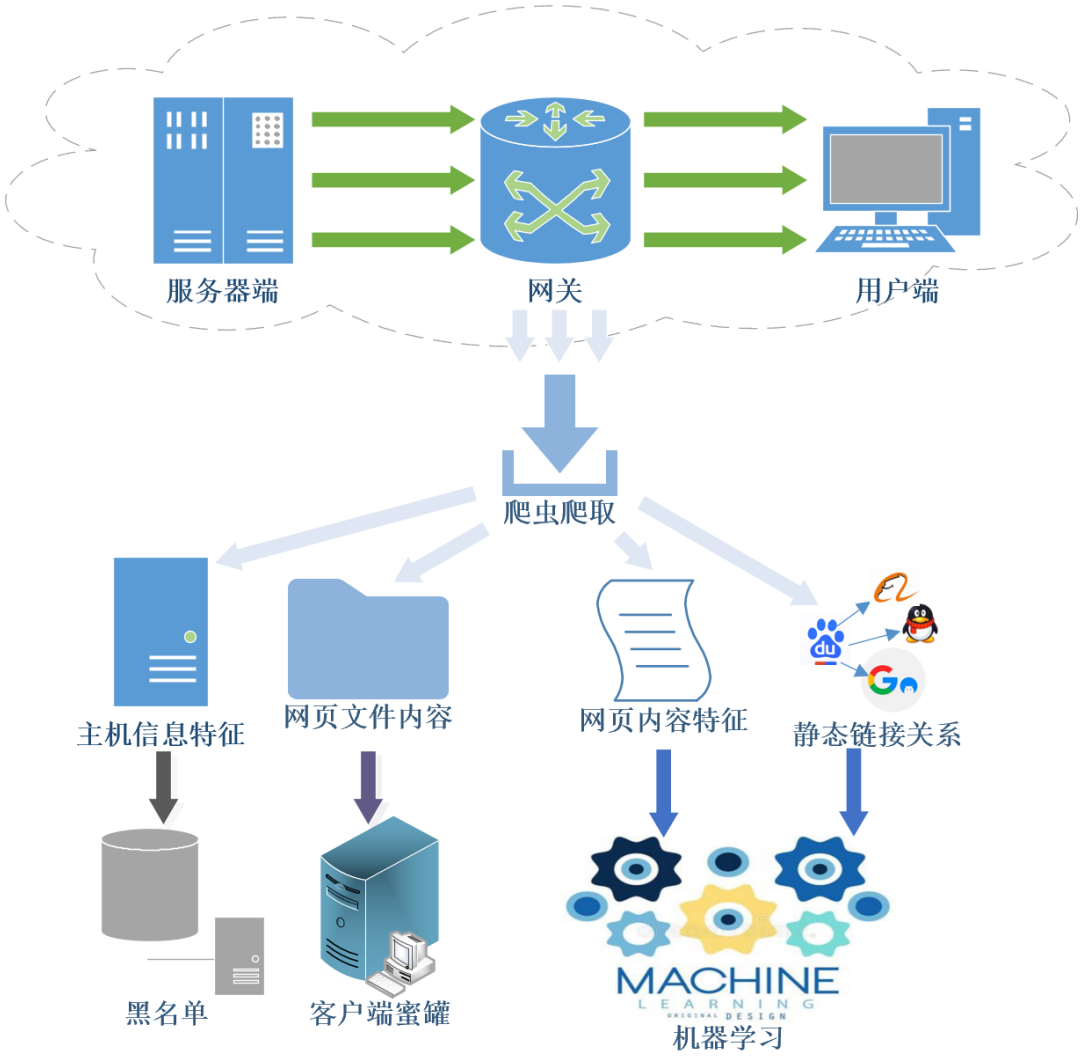

图 1.2.1 系统核心功能原理演示图

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：

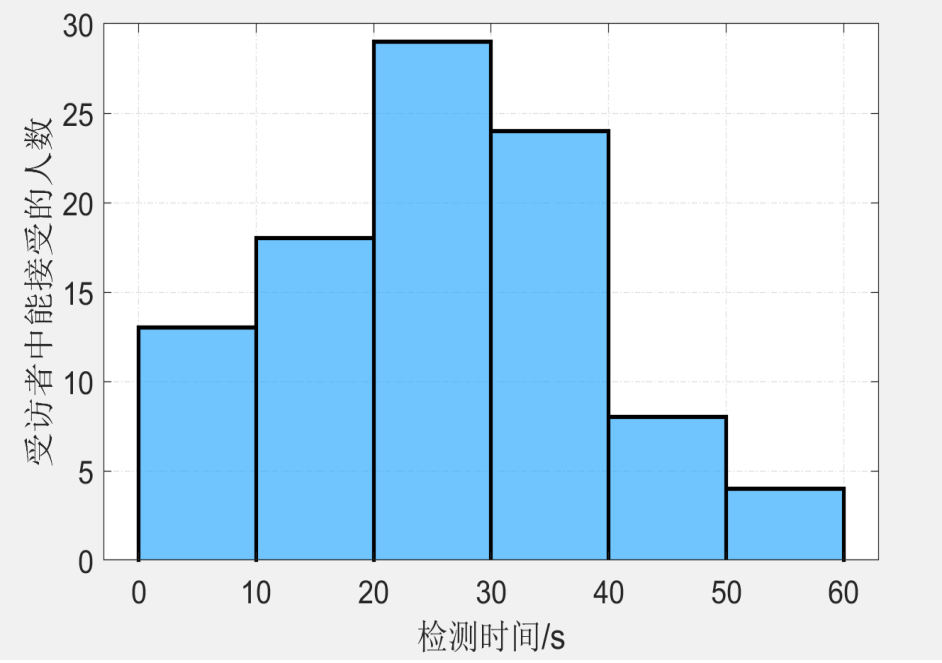

图 1.4.2.3

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。