## 第３９卷 第３期

２０１６年３月

计 算 机 学 报

ＣＨＩＮＥＳＥＪＯＵＲＮＡＬ ＯＦＣＯＭＰＵＴＥＲＳ

Ｖｏｌ．３９ Ｎｏ．３

Ｍａｒ．２０１６

恶意网页识别研究综述

沙泓州１），２），３） 刘庆云１），３） 柳厅文１），３） 周 舟１），３） 郭 莉１），３） 方滨兴２），３）

１）（中国科学院信息工程研究所 北京 １０００９３）

２）（北京邮电大学计算机学院 北京 １００８７６）

３）（信息内容安全技术国家工程实验室 北京 １０００９３）

## 摘 要 近年来，随着互联网的迅速发展以及网络业务的不断增长，恶意网页给人们的个人隐私和财产安全造成的威胁日趋严重．恶意网页识别技术作为抵御网络攻击的核心安全技术，可以帮助人们有效避免恶意网页引起的 安全威胁，确保网络安全．文中从理论分析和方法设计两方面介绍了恶意网页识别的最新研究成果．在理论分析层 面，从恶意网页的基本概念和形式化定义出发，对恶意网页识别的应用场景、基本框架及评价方法进行全面的归 纳，并总结了恶意网页识别的理论依据及性能评价指标．在方法设计层面，对具有影响力的恶意网页识别方法进行 了介绍和归类，对不同类别的识别方法进行了定性分析和横向比较．在总结恶意网页识别研究现状的基础上，从客 观环境的变化以及逃逸技术的升级两方面深入探讨了当前恶意网页识别面临的技术挑战．最后总结并展望了恶意 网页识别的未来发展方向．

关键词 恶意网页识别；网页分类；机器学习；逃逸技术

中图法分类号 ＴＰ３９３ ＤＯＩ号 １０．１１８９７／ＳＰ．Ｊ．１０１６．２０１６．００５２９

Ｓｕｒｖｅｙｏｎ ＭａｌｉｃｉｏｕｓＷｅｂｐａｇｅＤｅｔｅｃｔｉｏｎＲｅｓｅａｒｃｈ

ＳＨＡ Ｈｏｎｇ－Ｚｈｏｕ１），２），３） ＬＩＵ Ｑｉｎｇ－Ｙｕｎ１），３） ＬＩＵ Ｔｉｎｇ－Ｗｅｎ１），３） ＺＨＯＵ Ｚｈｏｕ１），３） ＧＵＯ Ｌｉ１），３） ＦＡＮＧＢｉｎ－Ｘｉｎｇ２），３）

１）（ＩｎｓｔｉｔｕｔｅｏｆＩｎｆｏｒｍａｔｉｏｎＥｎｇｉｎｅｅｒｉｎｇ，ＣｈｉｎｅｓｅＡｃａｄｅｍｙｏｆＳｃｉｅｎｃｅｓ，Ｂｅｉｊｉｎｇ １０００９３）

２）（ＳｃｈｏｏｌｏｆＣｏｍｐｕｔｅｒＳｃｉｅｎｃｅ，ＢｅｉｊｉｎｇＵｎｉｖｅｒｓｉｔｙｏｆＰｏｓｔｓａｎｄＴｅｌｅｃｏｍｍｕｎｉｃａｔｉｏｎｓ，Ｂｅｉｊｉｎｇ １００８７６）

３）（ＮａｔｉｏｎａｌＥｎｇｉｎｅｅｒｉｎｇＬａｂｏｒａｔｏｒｙｆｏｒＩｎｆｏｒｍａｔｉｏｎＳｅｃｕｒｉｔｙＴｅｃｈｎｏｌｏｇｙ，Ｂｅｉｊｉｎｇ １０００９３）

# Ａｂｓｔｒａｃｔ Ｉｎｒｅｃｅｎｔｙｅａｒｓ，ｗｉｔｈｔｈｅｒａｐｉｄｄｅｖｅｌｏｐｍｅｎｔｏｆＩｎｔｅｒｎｅｔａｎｄｔｈｅｉｎｃｒｅａｓｉｎｇｇｒｏｗｔｈｏｆ ｎｅｔｗｏｒｋｓｅｒｖｉｃｅｓａｎｄｓｅｃｕｒｉｔｙｎｅｅｄｓ，ｔｈｅｅｘｉｓｔｅｎｃｅｏｆｍａｌｉｃｉｏｕｓｗｅｂｐａｇｅｓｈａｖｅｂｅｃｏｍｅａ ｍｕｃｈ ｍｏｒｅｓｅｒｉｏｕｓｐｒｏｂｌｅｍｆｏｒｐｅｒｓｏｎａｌｐｒｉｖａｃｙａｎｄｐｒｏｐｅｒｔｙｓａｆｅｔｙ．Ａｓｏｎｅｏｆｔｈｅｋｅｙｔｅｃｈｎｏｌｏｇｉｅｓｔｏ ｒｅｓｉｓｔｎｅｔｗｏｒｋａｔｔａｃｋｓ，ｔｈｅｄｅｔｅｃｔｉｏｎｔｅｃｈｎｉｑｕｅｓｆｏｒ ｍａｌｉｃｉｏｕｓ ｗｅｂ ｐａｇｅｓｃａｎｅｆｆｅｃｔｉｖｅｌｙ ｈｅｌｐ ｐｅｏｐｌｅａｖｏｉｄｐｏｔｅｎｔｉａｌｓｅｃｕｒｉｔｙｔｈｒｅａｔｓａｎｄｔｈｕｓｅｎｓｕｒｅｔｈｅｎｅｔｗｏｒｋｓｅｃｕｒｉｔｙ．Ｉｎｔｈｉｓｐａｐｅｒ，ｗｅ ｄｅｓｃｒｉｂｅｔｈｅｌａｔｅｓｔｒｅｓｅａｒｃｈａｃｈｉｅｖｅｍｅｎｔｓｆｒｏｍｔｈｅｏｒｙｔｏｐｒａｃｔｉｃｅ．Ｉｔｓｔａｒｔｓｆｒｏｍｔｈｅｉｎｔｒｏｄｕｃｔｉｏｎ ｏｆｔｈｅｆｏｒｍａｌｄｅｆｉｎｉｔｉｏｎ ｏｆ ｍａｌｉｃｉｏｕｓ ｗｅｂ ｐａｇｅｓ，ａｎｄ ｆｏｌｌｏｗｅｄ ｂｙ ｃｏｎｃｌｕｄｉｎｇ ｔｈｅ ｄｅｔｅｃｔｉｏｎ ｔｅｃｈｎｉｑｕｅｓ’ａｐｐｌｉｃａｔｉｏｎｓｃｅｎａｒｉｏｓ，ｂａｓｉｃｆｒａｍｅｗｏｒｋａｎｄｅｖａｌｕａｔｉｏｎｐｒｉｎｃｉｐｌｅｓ．Ｔｈｅｎ，ｉｔｉｎｔｒｏｄｕｃｅｓ ｓｅｖｅｒａｌｔｙｐｉｃａｌｄｅｔｅｃｔｉｏｎｓｃｈｅｍｅｓ，ｃｌａｓｓｉｆｉｅｓｔｈｅｍ ｉｎｔｏｃａｔｅｇｏｒｉｅｓ，ａｎｄｆｉｎａｌｌｙ ｐｕｔｓｔｈｅｍ ｔｏａ ｈｏｒｉｚｏｎｔａｌｃｏｍｐａｒｉｓｏｎ．Ｂａｓｅｄｏｎｔｈｅｕｎｄｅｒｓｔａｎｄｉｎｇｏｆｔｈｅｒｅｓｅａｒｃｈｓｔａｔｕｓｉｎ ｍａｌｉｃｉｏｕｓｗｅｂｐａｇｅ ｄｅｔｅｃｔｉｏｎｓｃｈｅｍｅｓ，ｔｈｉｓｐａｐｅｒｐｒｅｓｅｎｔｓａｎｉｎ－ｄｅｐｔｈｄｉｓｃｕｓｓｉｏｎｏｆｔｈｅｃｕｒｒｅｎｔｃｈａｌｌｅｎｇｅｓｉｎ ｗｈｉｃｈ

收稿日期：２０１５－０１－０７；最终修改稿收到日期：２０１５－０５－２１．本课题得到中国科学院战略性先导科技专项（ＸＤＡ０６０３０２００）、国家科技支撑计划（２０１２ＢＡＨ４６Ｂ０２）和国家自然科学基金项目（６１４０２４７４）资助．沙泓州，男，１９８８年生，博士研究生，中国计算机学会（ＣＣＦ）学生会员，主要研究方向为信息安全、网络安全．Ｅ－ｍａｉｌ：ｓｈａｈｏｎｇｚｈｏｕ＠ｆｏｘｍａｉｌ．ｃｏｍ．刘庆云（通信作者），男，１９８０ 年生，博士，高级工程师，主要研究方向为信息安全、网络安全．Ｅ－ｍａｉｌ：ｌｉｕｑｉｎｇｙｕｎ＠ｉｉｅ．ａｃ．ｃｎ．柳厅文，男，１９８６年生，博士，助理研究员，主要研究方向为大数据安全分析、算法分析与设计．周 舟，男，１９８３年生，博士，高级工程师，主要研究方向为网络安全、高性能网络．郭 莉，女，１９６９ 年生，高级工程师，主要研究方向为信息安全、网络安全、数据流处理．方滨兴，男，１９６０年生，博士，教授，博士生导师，中国工程院院士，主要研究领域为网络安全、信息内容安全．

## ５３０

计 算 机 学 报 ２０１６年

ｐｅｏｐｌｅｈａｖｅｔｏｆａｃｅ，ｉｎｃｌｕｄｉｎｇｂｏｔｈｄｙｎａｍｉｃａｌｃｈａｎｇｅｓｏｆｔｈｅｏｂｊｅｃｔｉｖｅｅｎｖｉｒｏｎｍｅｎｔｓａｎｄｕｐｇｒａｄｅｓ ｏｆｔｈｅｅｓｃａｐｅｔｅｃｈｎｉｑｕｅｓ．Ｆｉｎａｌｌｙ，ｉｔｌｏｏｋｓｉｎｔｏｔｈｅｆｕｔｕｒｅｏｆｔｈｉｓｆｉｅｌｄ．

Ｋｅｙｗｏｒｄｓ ｍａｌｉｃｉｏｕｓ ｗｅｂ ｐａｇｅｄｅｔｅｃｔｉｏｎ；ｗｅｂ ｐａｇｅｃｌａｓｓｉｆｉｃａｔｉｏｎ；ｍａｃｈｉｎｅｌｅａｒｎｉｎｇ；ｅｓｃａｐｅ ｔｅｃｈｎｏｌｏｇｙ

１ 引 言

近年来，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．与此同时，便捷的网络服务也吸引了网络攻击者们通过钓鱼网站［１］、垃 圾广告［２］和恶意软件［３］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．这些网页因此被称为恶意网页．卡巴斯基报告①指出，２０１２年，恶意网页在８７．３６％ 的网络攻击中出现，并已成为黑客谋求不法经济利益的重要工具．此 外，Ｇｏｏｇｌｅ 的研 究［４］ 指 出，其 搜 索 结 果 中

１．３％ 的页面为挂马网页．因此，如何有效地识别恶

意网页已经成为亟待解决的网络安全问题之一． 针对恶意网页识别问题，研 究人员提出很多

识别技术和解决方案．Ｍａ等人［５－７］依据 ＤＮＳ 信息、

ＷＨＯＩＳ信息以及 ＵＲＬ 的语法特征，采用机器学习算法对恶意 ＵＲＬ 进行识别．Ｃａｎａｌｉ等人［８］在此基础上增加了Ｊａｖａｓｃｒｉｐｔ和 ＨＴＭＬ 特征，从而能够对网页内容进行检测，提升了恶意网页识别准确率．此外，Ｔｈｏｍａｓ等人［９］应用蜜罐技术分析浏览器响应动作（例如Ｊａｖａｓｃｒｉｐｔ事件及弹出新窗口等），以此判定当前访问页面是否为恶意网页．而一些浏览器出于快速响应需要采用内置恶意网址列表的方法为用户提供轻量级的实时恶意网页识别服务（例 如

ＩｎｔｅｒｎｅｔＥｘｐｌｏｒｅｒ 浏览器的 ＳｍａｒｔＳｃｒｅｅｎ 筛选器②及谷歌浏览器的 ＳａｆｅＢｒｏｗｓｉｎｇ③ 等）．随着网络服务日益普及，围绕恶意网页的攻防博弈也在持续进行，攻击者不断采用一些新的技术手段（例如自动生成域名［１０］及网页隐匿技术［１１］等）来增强恶意网页的隐蔽性，提高攻击效率；而防御方安全研究人员深入研究恶意网页识别技术，不断提出相应的检测手段

和防范措施．

从公开发表的科研论文和资料来看，国内外对恶

意网页识别的相关研究还不够全面和深入．２０１２ 年，张慧琳等人［１２］从网页木马的机理出发，从挂马检测、

特征分析、防范技术等方面对网页木马的研究进行了分析和总结，但缺少对各类检测方法的横向比较和讨论；２０１３ 年，诸葛建伟等人［１３］对识别恶意网页常用的蜜罐技术的研究发展和应用情况进行了综述，但没有涉及其它恶意网页识别技术．此外，２０１３年，

Ｍａｈｍｏｕｄ等人［１］对钓鱼网站的离线识别方法进行

了详细地分析和总结，但缺少对其他恶意网页以及在线识别方法的分析．

因此，本文尝试对恶意网页识别的思路、方法、

技术进行全面的归纳和总结，介绍恶意网页识别系统的整体框架和应用场景，详细分析恶意网页识别的研究现状并探讨防御方目前面临的挑战，为进一步研究作参考．

本文第２节介绍恶意网页的基本概念与形式化

定义；第３节对恶意网页识别技术进行概述；第４节和第５节对恶意网页识别的研究进展和挑战进行总结和分析；第６节对未来工作进行展望并对全文工作进行总结．

２ 恶意网页基本概念与形式化定义

２．１ 恶意网页基本概念

目前，学术界对恶意网页尚无一个明确的、统一的定义．Ｇｏｏｇｌｅ③ 将恶意网页限定为一种不安全的网站，发生的场景可以是恶意软件自动下载［１４］，网页弹窗［１］诱骗用户输入自己的用户名和密码等．而

Ｅｓｈｅｔｅ等人［１５］将恶意网页定义为一类利用漏洞对

一次性的访问行为发起攻击的网页．此外，百度百科上④将恶意网站定义为故意在计算机系统上执行恶意任务的病毒、蠕虫和特洛伊木马的非法网站，并指出它们的共同特征是采用网页形式让人们正常浏览

① Ｋａｐａｅｒｓｋｙ． Ｋａｐａｅｒｓｋｙ ｓｅｃｕｒｉｔｙ ｂｕｌｌｅｔｉｎ． ｈｔｔｐ：／／ｗｗｗ． ｓｅｃｕｒｅｌｉｓｔ．ｃｏｍ／ｅｎ／ａｎａｌｙｓｉｓ／２０４７９２２５５／Ｋａｐｅｒｓｋｙ，２０１２．１０．１０

② Ｉｎｔｅｒｎｅｔ Ｅｘｐｌｏｒｅｒ．ＳｍａｒｔＳｃｒｅｅｎ ｆｉｌｔｅｒ．ｈｔｔｐ：／／ｗｉｎｄｏｗｓ．

ｍｉｃｒｏｓｏｆｔ．ｃｏｍ／ｚｈ－ＣＮ／ｉｎｔｅｒｎｅｔ－ｅｘｐｌｏｒｅｒ／ｕｓｅ－ｓｍａｒｔｓｃｒｅｅｎ－ｆｉｌｔｅｒ

＃ｉｅ＝ｉｅ－９，２０１４．１２．１２

③ Ｇ．Ｔ．Ｒｅｐｏｒｔ．Ｍａｋｉｎｇｔｈｅ Ｗｅｂｓａｆｅｒ．ｈｔｔｐ：／／ｗｗｗ．ｇｏｏｇｌｅ． ｃｏｍ／ｔｒａｎｓｐａｒｅｎｃｙｒｅｐｏｒｔ／ｓａｆｅｂｒｏｗｓｉｎｇ／？ｈｌ＝ｅｎ，２０１４．１２．３０

④ Ｂａｉｄｕ．Ｄｅｆｉｎｉｔｉｏｎ ｏｆ ｍａｌｉｃｉｏｕｓ Ｗｅｂ ｓｉｔｅｓ．ｈｔｔｐ：／／ｂａｉｋｅ．

ｂａｉｄｕ．ｃｏｍ／ｖｉｅｗ／２３８２１１９．ｈｔｍ，２０１４．１２．３１

# 页面内容，同时非法获取电脑里的各种数据．该定义指出了“访问页面时执行恶意行为”和“非法窃取用户数据”两个关键点．

综上所述，恶意网页是一类以网页木马、钓鱼网

站为代表的网页．不同于正常网页，恶意网页通过伪装成合法网站或在页面中嵌入恶意脚本等方式，在用户访问时对其网络安全构成威胁．

因此，本文将恶意网页定义为以网页形式出现，

以访问时窃取用户隐私、安装恶意程序或执行恶意代码等恶意行为为目的的网页集合．

２．２ 恶意网页识别形式化定义

恶意网页识别问题的本质是一个二分类问题，可以形式化定义如下：设 Ｗ 表示网页样本集合：

Ｗ ＝｛ｗ１，ｗ２，…，ｗｉ，…，ｗｎ｝，其中ｎ 为网页数量，ｗｉ

为第ｉ个网页．Ｃ 表示网页类标号集合：Ｃ＝｛ｃｌ，ｃｍ ｝，其中ｃｌ表示良性网页，ｃｍ 表示恶意网页．则目标函数为从网页样本集合到类标号集合的映射函数：

*Φ*（ｗｉ，ｃｊ）：Ｗ ×Ｃ→ ｛０，１｝．其中，１ｉｎ，ｊ∈｛ｌ，ｍ｝．

*Φ*（ｗｉ，ｃｊ）是一个二分类函数．因为恶意网页识别只识别网页 ｗｉ是否属于恶意网页集合，所以，该目标函数可以简化为*Φ*（ｗｉ，ｃｊ）：Ｗ →｛０，１｝．

３ 恶意网页识别技术概述

本节依据已有的恶意网页识别系统和相关研究工作，首先分析恶意网页识别的主要应用场景，然后对恶意网页识别系统的基本框架进行归纳，并总结相应的评价指标．

３．１ 恶意网页识别的应用场景

恶意网页规模的不断壮大，在给人们的个人隐私和财产安全带来威胁的同时，客观上也为恶意网页识别技术创造了庞大的安全市场和广泛的应用需求．下面将从恶意网页的攻击形式以及检测位置等方面分析恶意网页识别技术的应用场景．

按照攻击形式来分，恶意网页主要分为钓鱼网

页［１６］和恶意软件下载［１２］两大类．其中，依据中国互联网协会的定义⑤，恶意软件作为一个集合名词，指代在计算机系统上执行恶意任务［１７］的蠕虫、病毒、网页木马、间谍软件等．由于采用的攻击形式不同， 这些恶意网页对访问者构成了不同类型的安全威

者的计算机系统安全构成威胁．

按照检测位置来分，识别恶意网页的位置可以分为３类：服务器端（例如搜索引擎［１８］及社交网站［１９］等），客户端（例如提供安全服务的浏览器插件② 、杀毒软件⑥等）以及网关端（例如安全网关设备［２０］等）．其中，大多数恶意网页识别系统［１８－１９］在服务器端和客户端进行识别和检测．

３．２ 恶意网页识别系统基本框架

为了准确识别数量众多、种类多样的恶意网页， 恶意网页识别系统应当同时具备可用性与可扩展

性．由于目前已有的恶意网页识别系统大都面向某一类特定应用，因此系统结构和实现方式存在一些差异．下面结合恶意网页识别的目标和一般规律，提炼出恶意网页识别系统的基本框架，如图１所示，该框架主要分为３个部分．

（１）网页采集．负责对互联网上的网页进行收

集、去重和过滤．其中，按照网页收集方式，一般可分 为主动和被动两种．主动收集，主要是运用网络爬虫 技术，从互联网中定向抓取网页集合．而被动收集， 主要是在网关或客户端蜜罐中，对经过的访问流量进行采集．而流量过滤是根据网页的相关信息（例如 网页后缀及网页类型等），对明显不属于恶意网页的 部分进行过滤，以提升恶 意网页识别系统的运行效率．

（２）特征抽取．特征抽取是指不同的识别方法

从识别不同种类恶意网页的实际需求出发，依据网页自身特点，抽取信息作为识别恶意网页依据的过程．这些特征包括但不限于 ＵＲＬ 词汇特征、主机信息特征、网页内容特征、ＵＲＬ（ＤＮＳ）黑名单、链接关系以及跳转关系等．如图２所示，针对不同类别的恶意网页，学者们从不同角度提出了很多识别恶意网页的特征．常用的识别特征按照其来源的不同可以分为静态特征和动态特征两类．

静态特征主要来源于网页静态信息，其种类繁

多，但抽取过程相对简单．常见的静态特征主要包括主机信息（例如 ＷＨＯＩＳ信息［２１］及 ＤＮＳ信息［２２］等），

ＵＲＬ 信息（例如词袋特征［２３］及 ＵＲＬ 长度［２４］等）和

# 网页内容（例如 ＨＴＭＬ Ｔａｇ 信息［２５］、Ｊａｖａｓｃｒｉｐｔ代码［２６］、网 页漏洞信息［２７］及 链接关系［２８］）等．而 动态特征主要来自于网页动态行为，其种类较少，但抽

胁．一些恶意网页常用于窃取用户的个人隐私信息

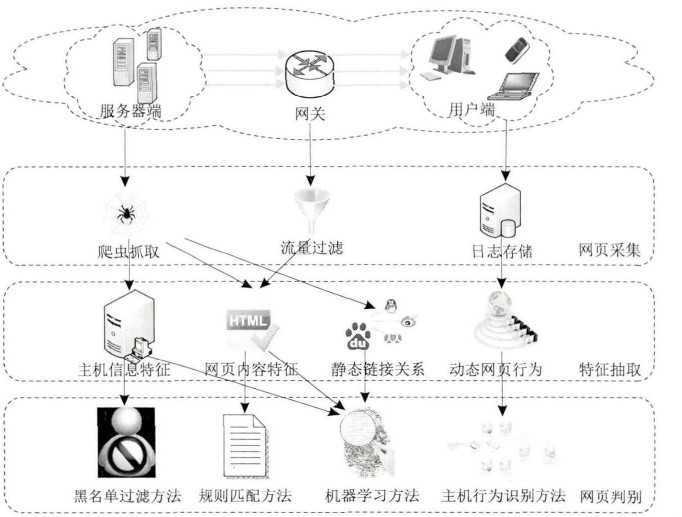
（例如攻击者常利用钓鱼网页窃取用户的银行帐号

及密码信息等），而另一些恶意网页则通过下载和执行恶意程序或脚本（如病毒、木马、蠕虫等），对访问

⑤ Ｄｅｆｉｎｉｔｉｏｎｏｆ ｍａｌｗａｒｅ．ｈｔｔｐ：／／ｔｅｃｈ．ｓｉｎａ．ｃｏｍ．ｃｎ／ｉ／２００６－ １１－２２／２０３１１２５１６４５．ｓｈｔｍｌ，２０１４．１２．３０

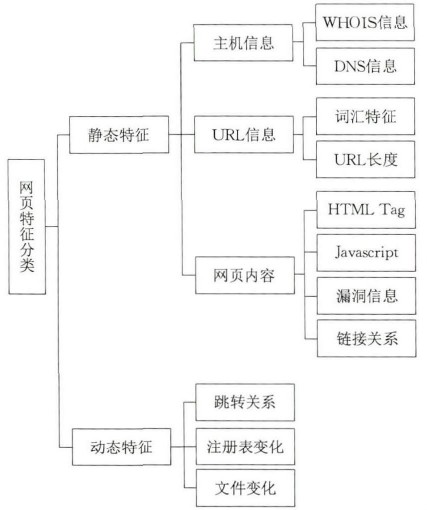
⑥ ＭｃＡｆｅｅ： Ｍｃａｆｅｅ ｓｉｔｅ ａｄｖｉｓｏｒ．ｈｔｔｐ：／／ｗｗｗ．ｓｉｔｅａｄｖｉｓｏｒ．

ｃｏｍ／，２０１４



## 图 ２ 识别恶意网页的特征分类

图 １ 恶意网页识别的基本框架

黑名单过滤法、规则匹配法、机器学习方法及基于交互式主机行为的识别方法．本文将在第４ 节对此作详细介绍．

３．３ 恶意网页识别评价指标

用于识别恶意网页的评价指标，通常可以分为以下两类：功能指标和性能指标．其中，功能指标主 要用于对恶意网页的识别效果进行评价；而性能指 标主要对恶意网页的识别效率进行评价．为了方便 对功能指标进行描述，表１ 展示了常用于衡量分类准确性的混淆矩阵［３４］．其中，Ｎ 表示恶意网页被正确划分的数量，ＮＬ→ Ｍ′ 表示良性网页被错误划分为恶意网页的数量（常被称为误报），ＮＭ →Ｌ′ 表示恶意网页被错划分为良性网页的数量（常被称为漏报），ＮＬ→Ｌ′表示良性网页被正确划分的数量．其中，漏报和误报是恶意网页识别中可能出现的两种错误情况．举例来说，如有２００ 个网页，分为良性网页

Ｍ → Ｍ′

和恶意网页两类，各１００个．表１展示了由恶意网页

取过程相对复杂．常见的动态特征主要包括浏览器行为［２９］、网页跳转关系［３０］、注册表及文件夹变化情况［１３］等．这些特征往往需要人们对可疑的网页进行长时间的深入分析才能获得．同时，在使用动态特征的过程中，往 往需要结合蜜网技术⑦ 和虚拟化技术［３１］辅助识别恶意网页．此外，一些研究从 ＨＴＴＰ会话［３２］、搜索引擎提供的相似网页［３３］出发，为识别恶意网页提供新思路．

（３）网页判别．目前常用的网页判别方法包括

识别系统处理得到的一种可能情况．在表１中，误报

２０个，漏报１０个．

表 １ 分类混淆矩阵

认定为恶意网页（Ｍ′） 认定为良性网页（Ｌ′）

恶意网页（Ｍ） ＮＭ → Ｍ′（如９０） ＮＭ →Ｌ′（如１０）

良性网页（Ｌ） ＮＬ→ Ｍ′（如２０） ＮＬ→Ｌ′（如８０）

⑦ Ｔｈｅ Ｈｏｎｅｙｎｅｔ Ｐｒｏｊｅｃｔ．Ｃａｐｔｕｒｅ－ＨＰＣ．ｈｔｔｐ：／／ｐｒｏｊｅｃｔｓ． ｈｏｎｅｙｎｅｔ．ｏｒｇ／ｃａｐｔｕｒｅ－ｈｐｃ，２０１４．１０．２０

# 依据混淆矩阵，恶意网页识别的功能指标主要有： 真正类率（ＴｒｕｅＰｏｓｉｔｉｖｅＲａｔｅ，ＴＰＲ）．由式（１）

计算，表示在所有恶意实例中被检测出的恶意实例

的比例．

假正类率（ＦａｌｓｅＰｏｓｉｔｉｖｅＲａｔｅ，ＦＰＲ），也称误报率．由式（２）计算，表示在所有良性实例中被错误检测为恶意实例的良性实例的比例．

真负类率（ＴｒｕｅＮｅｇａｔｉｖｅＲａｔｅ，ＴＮＲ）．由式（３）计算，表示在所有良性实例中被正确检测的良性实例的比例．

假负类率（ＦａｌｓｅＮｅｇａｔｉｖｅＲａｔｅ，ＦＮＲ），也称漏报率．由式（４）计算，表示在所有恶意实例中被错误检测为良性实例的恶意实例的比例．

精确度（Ｐｒｅｃｉｓｉｏｎ）．由式（５）计算，表示在所有被检测出的恶意实例中正确的恶意实例的比例．

召回率（Ｒｅｃａｌｌ），也称查全率．等价于 ＴＰＲ，由式（６）计算．

Ｆ－ｍｅａｓｕｒｅ是准确率和查全率的加权调和平均，可由式（７）计算得到，其中*β*为参数．当参数*β*＝１时，就是最常见的Ｆ１－ｍｅａｓｕｒｅ，可由式（８）计算得到．

准确度（Ａｃｃｕｒａｃｙ）．由式（９）计算得到，表示在所

有实例中正确检测出的良性实例和恶意实例的比例．

对于一些实时或者近实时系统［３５－３６］来说，快速识别和发现新的恶意网页十分重要．因此，Ｉｎｖｅｒｎｉｚｚｉ等人［３３］以原有的恶意网页作为基础，寻找和识别新的恶意网页．此时，原有的评价指标不再适用．于是， 文献［３３］提出了一类新的评价指标：毒性（ｔｏｘｉｃｉｔｙ）和扩展度（ｅｘｐａｎｓｉｏｎ）．毒性是指所分析网页中真正恶意网页所占的比例．毒性越高，识别单个网页所需 要消耗的计算资源越少．扩展度是指从一个恶意网页出发可以发现的新恶意网页数目．扩展度越高，意 味着原有恶意网页的利用越充分．对同一恶意网页识别系统而言，毒性和扩展度一般存在着此消彼长的关系，即毒性越高，扩展度越低．

常用的性能指标主要包括处理一个网页所需要

的时间、单位时间内网页吞吐量等．

４ 恶意网页识别研究进展

目前，恶意网页识别的方法主要包括４类：基于黑名单技术的识别方法、基于启发式规则的识别方法、基于机器学习的识别方法以及基于交互式主机行为的识别方法．

４．１ 基于黑名单技术的识别方法

ＴＰ ＝ ＮＭ → Ｍ

ＮＭ → Ｍ ＋ＮＭ →Ｌ

ＦＰ ＝ ＮＬ→ Ｍ

ＮＬ→Ｌ ＋ＮＬ→ Ｍ

ＴＮ ＝ ＮＬ→Ｌ

ＮＬ→Ｌ ＋ＮＬ→ Ｍ

Ｎ

（１）

# （２）

（３）

黑名单是一份包含恶意网页 ＵＲＬ、ＩＰ 地址或者关键词信息的列表．通过使用黑名单技术，人们可以准确识别已被确认的恶意网页，从而降低误报率ＦＰＲ．根据包含信息种类的不同，黑名单可以细分为

ＵＲＬ 黑名单、ＩＰ 地址黑名单以及 ＤＮＳ 黑名单等．黑名单技术实现简单，使用方便，因而广泛应用于

ＦＮ ＝

ＮＭ →

Ｍ →Ｌ

Ｍ ＋ＮＭ →Ｌ

（４）

# ＧｏｏｇｌｅＳａｆｅＢｒｏｗｓｉｎｇ③ 、ＭａｌｗａｒｅＤｏｍａｉｎＬｉｓｔ⑧ 及

Ｐ ＝ ＮＭ → Ｍ

ＮＬ→ Ｍ ＋ＮＭ → Ｍ

（５）

# ＰｈｉｓｈＴａｎｋ⑨ 等实际项目和系统中．在实际应用中，黑名单技术常常需要与人工检查、蜜网技术［３７］等其

Ｒ ＝ ＴＰ （６）

他技术配合使用．

Ｆ*β* ＝

（２＋１）ＰＲ

# ２Ｐ ＋Ｒ

*β*

*β*

（７）

# 以 ＧｏｏｇｌｅＳａｆｅＢｒｏｗｓｉｎｇ ＡＰＩ③ 为例，它 根据

Ｇｏｏｇｌｅ提供的持续更新的 ＵＲＬ 列表，允许用户检

Ｆ１ ＝ ２ＰＲ

# Ｐ ＋Ｒ

ＮＬ→Ｌ ＋ＮＭ → Ｍ

（８）

# 查特定 ＵＲＬ 是否存在于这个列表上，以判断其是否为网络钓鱼或恶意软件．而 基于域名的黑名单

ＡＣＣ ＝

Ｎ

Ｌ→Ｌ

＋ＮＬ→ Ｍ

＋ＮＭ →Ｌ

＋ＮＭ → Ｍ

（９）

# （ＭａｌｗａｒｅＤｏｍａｉｎＬｉｓｔ⑧ ）则主要依据域名或者ＩＰ

ｎ Ｎ 地址信息识别和过滤对特定网站或网址的访问行

∑ｉ＝１

Ｍ →Ｌ

# （ ）

ＣＥＲｎ ＝

ｎ

∑ｉ＝１（ＮＭ →Ｌ ＋ＮＭ → Ｍ

# ） １０ 为．此外，ＰｈｉｓｈＴａｎｋ⑨ 为人们提供了一个自愿提交

和共享钓鱼网页网址的开放平台．人 们可以依据

此外，在实时恶意网页识别系统中，累计错误率

（ＣｕｍｕｌａｔｉｖｅＥｒｒｏｒＲａｔｅ，ＣＥＲ）是十分重要的功能评价指标之一．累计错误率是一个随时间变化的累计值，第ｎ 天的累计错误率可由式（１０）计算得到．

⑧ Ｍａｌｚｉｌｌａ．Ｍａｌｗａｒｅ Ｄｏｍａｉｎ Ｌｉｓｔ．ｈｔｔｐ：／／ｗｗｗ．ｍａｌｗａｒｅ－ ｄｏｍａｉｎｌｉｓｔ．ｃｏｍ，２０１４．１０．２０

⑨ ＯｐｅｎＤＮＳ． ＰｈｉｓｈＴａｎｋ． ｈｔｔｐ：／／ｗｗｗ．ｐｈｉｓｈｔａｎｋ．ｃｏｍ， ２０１４．１０．２０

# ＰｈｉｓｈＴａｎｋ 提供的列表主动过滤钓鱼网址，从而保障网络安全．

然而，黑名单仅能识别已经发现的恶意网页，而

不能正确识别之前未出现的恶意网页，从而容易引起漏判．为了改善漏判情况，Ｐｒａｋａｓｈ等人［３８］针对黑名单技术提出了一种改进方法 ＰｈｉｓｈＮｅｔ．ＰｈｉｓｈＮｅｔ将已经发现的钓鱼 ＵＲＬ 作为先验知识，通过 ＵＲＬ 分解和相似性计算来识别和发现新的钓鱼网页．通过这种方式，ＰｈｉｓｈＮｅｔ扩展了黑名单的使用范围，有助于识别部分未出现的恶意网页，但它的识别能力依赖于原有黑名单集合的规模，并存在时间开销随黑名单规模扩大而线性增长的缺点．

除了上述漏判和时间开销大的问题，黑名单还

存在更新时效性低的缺点．在使用黑名单技术的过程中，当发现疑似恶意网页时，首先由综合检查技术进行分析确认，然后依据实际更新策略进行分发和部署．综合检查技术的水平决定了新恶意网址的确认时间，而更新策略决定了每次更新内容的实际生效时间．由于网络和计算资源的限制等原因，每次更新的时效性难以保证．以钓鱼网站为例，根据 Ｓｈｅｎｇ等人［３９］的研究，约有６３％的网络钓鱼行为在最初的

２ｈ 内就结束了，而４７％ ～８３％ 的钓鱼网址在发现

１２ｈ后才能录入黑名单．由此可见，黑名单更新时效性低的缺点将在很大程度上限制黑名单技术的实际使用效果．

４．２ 基于启发式规则的识别方法

为了克服黑名单机制所产生的漏判等缺点，研究人员设计并实现了基于启发式规则的恶意网页识别方法．这类方法的工作原理是依据恶意网页之间存在的相似性设计和实现启发式规则，进而发现和识别恶意网页．不同于黑名单依靠精确匹配完成恶意网页识别，基于启发式规则的方法不需要提前了解恶意网页的网址等信息，就可以依据现有规则识别部分未出现（未识别）的恶意网页．因此，它在一些主流浏览器上（包括火狐浏览器、ＩＥ 浏览器等）得到广泛应用，并且常以浏览器安全插件的形式出现．

２００４年，Ｃｈｏｕ等人［４０］开发出一套浏览器插件 ＳｐｏｏｆＧｕａｒｄ．ＳｐｏｏｆＧｕａｒｄ 部署在客户端，它的工作原理是：首先依据钓鱼网页常见情况建立启发式规则，从而对 ＨＴＭＬ 网页及其 ＵＲＬ（包括用户输入信息、链接关系及可疑的网址信息及图片信息等）进行检测．

２００７年，Ｚｈａｎｇ等人［４１］研发出一套针对ＩＥ 的

工具条 Ｃａｎｔｉｎａ，通过分析网页的词频－逆向文档频率（ＴＦ－ＩＤＦ），搜索返回结果及其他统计信息（例如网页中是否包含特殊字符－或＠及 ＵＲＬ 中点的个数等），建立启发式规则以判别当前网页是否是恶意网页．

基于启发式规则的方法往往假设对于某些恶意

网页，其统计特征（例如链接关系、网页内容是否包含关键词等）是唯一的，可以作为规则对恶意网页和良性网页进行区分．但是，对于大规模网页分类而言，简单的特征统计和启发式规则方法已经无法满足需求，主要体现在以下两个方面：

（１）误报率高．由于采用启发式规则的模糊匹配技术，这类方法将大大提升良性网页的误判概率． 因此，相较于黑名单方法，启发式规则的识别方法误 报率较高．

（２）规则更新难，依赖于领域知识．由于启发式规则是通过对已有恶意网页的特征统计或人工总结

得到的，因此这些规则依赖于对应的领域知识，因此更新困难．

４．３ 基于机器学习的识别方法

针对基于启发式规则识别方法存在的误报率高和规则更新难的问题，研究人员进一步提出了更加系统的基于机器学习的识别方法．

这类方法首先将恶意网页识别看作是一个文本

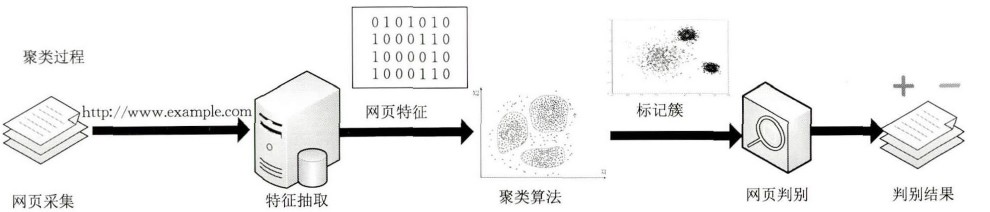
分类或聚类的问题，然后运用相应的机器学习算法

（例如ｋ－ｍｅａｎｓ、ＤＢＳＣＡＮ、ｋ－ＮＮ、Ｃ４．５ 及ＳＶＭ 等）进行识别．目前，用于恶意网页识别的机器学习方法 主要包括无监督方法和有监督方法．

４．３．１ 无监督机器学习方法

无监督机器学习方法又称聚类方法．这类方法首先将 ＵＲＬ 数据集划分为若干簇，使得同一簇的数据对象之间相似度较高，而不同簇的数据对象之间的相似度较低．然后，通过构造和标记数据集中的簇来区分恶意网页和良性网页．其具体分类过程如图３所示，主要由特征提取、聚类、簇标记和网页判别等步骤组成．

２０１０年，Ｌｉｕ 等人［４２］以链接关系、关键词排序关系、文本相似性关系、层次相似性关系等作为统计特征，利用无监督学习算法 ＤＢＳＣＡＮ［４３］（Ｄｅｎｓｉｔｙ－ ＢａｓｅｄＳｐａｔｉａｌＣｌｕｓｔｅｒｉｎｇｏｆＡｐｐｌｉｃａｔｉｏｎｓｗｉｔｈＮｏｉｓｅ）对钓鱼网页的攻击目标进行识别．实验结果表明，该方法可以识别９１．４４％ 钓鱼网页的攻击目标，并将误报率控制在３．４％以内．



４．３．２ 有监督机器学习方法

## 图 ３ 基于聚类算法的恶意网页识别过程

给定类别中的一个．分类算法的工作流程如图４ 所

有监督机器学习方法又称分类方法．研究人员通过引入网页信誉库的方式构造 ＵＲＬ 标注集，从而利用现有的分类算法为识别恶意网页提供了一类新思路．具体来说，他们根据已标记样本的特点构造分类规则或分类器，进而将未知类型的样本映射到

示，它主要包括两个步骤：

（１）训练．根据提供的训练数据集及特征集合构造一个分类模型．

（２）预测．依据训练过程产生的分类模型对未知样本的类别进行预测．

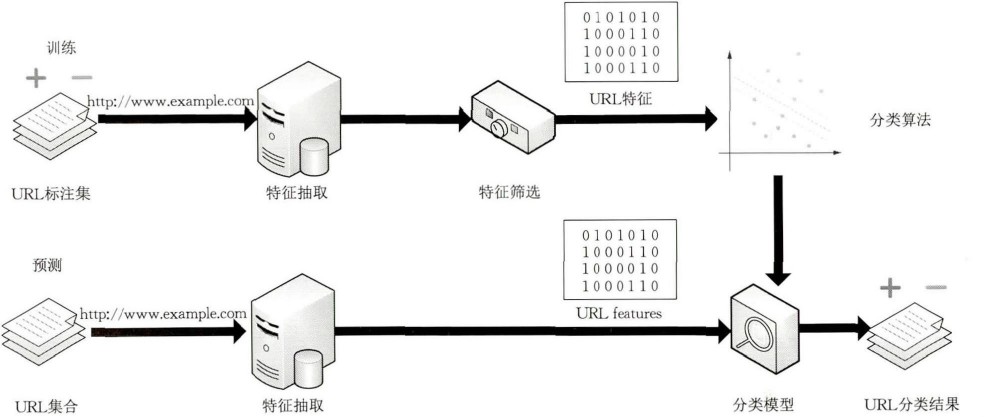


图 ４ 基于分类算法的恶意网页识别过程

其中用于训练和测试的数据需要具有相同的特征集．此外，训练集应当标记好网页所属的类别，并对特征向量进行归一化，其标准格式如表２所示．其中，Ｍ 代表恶意网页，Ｌ 代表良性网页，归一化后每个特征的取值范围一般为［０，１］．

表 ２ 网页特征集合示例

序号 类别 特征集示例

是否在黑名单 ＴＴＬ值 ＵＲＬ长度

１ Ｍ １ ０．２ ０．５

２ Ｌ ０ ０．５ ０．１

３ Ｌ ０ ０．３ ０．３

４ Ｍ ０ １．０ ０．８

５ Ｌ ０ ０．８ ０．２

# ３．２节详细介绍了机器学习算法所使用的特征集．本节将侧重于从分类模型选择方面介绍分类算法在恶意网页识别中的应用情况．分类算法按照实

时性不同分为离线分类算法和在线分类算法．其中， 常用的离线分类算法主要有决策树、贝叶斯、支持向量机（ＳｕｐｐｏｒｔＶｅｃｔｏｒＭａｃｈｉｎｅ，ＳＶＭ）、逻辑回归

（Ｌｏｇｉｓｔｉｃ Ｒｅｇｒｅｓｓｉｏｎ，ＬＲ）等．而 在线分类算法主要包括 ＰＡ（Ｐａｓｓｉｖｅ Ａｇｇｒｅｓｓｉｖｅ）、ＣＷ （Ｃｏｎｆｉｄｅｎｃｅ

Ｗｅｉｇｈｔｅｄ）算法以及ＡＲＯＷ（ＡｄａｐｔｉｖｅＲｅｇｕｌａｒｉｚａｔｉｏｎ

ｏｆＷｅｉｇｈｔｓ）等．

贝叶斯分类算法首先假设网页的每个特征之间相互独立．在已知网页特征向量的先验概率的情况下，它依据贝叶斯公式对其后验概率进行计算．通过比较不同网页的后验概率和预先设定好的阈值来判别网页是否属于恶意网页．Ｍａ 等人［５］使用朴素贝叶斯分类器在多个公开数据集上进行检测，其恶意网页识别的准确率在９４％ 以上．贝叶斯分类算法相对简单，可解释性强，分类速率快．它的局限性在于假设条件较强，在恶意网页识别领域，特征之间往往

并不独立．

和贝叶斯分类算法不同，支持向量机利用非线

性变换和结构风险最小化原则来提高分类器泛化能力，具有良好的分类准确率和稳定性．此外，它无需 满足特征相互独立的假设，并且在统计样本量很少的情况下也能获得较高的准确率．基于ＳＶＭ 算法，

Ｈｕａｎｇ等人［４４］提出了一套新的钓鱼网页识别系统，

并达到了９９％ 的识别准确率．ＳＶＭ 方法［４５］存在的主要缺陷在于它的可解释性较差，在高特征维度的情况下分类速率较低．此外，ＳＶＭ 方法的分类结果对训练集实际分布情况以及参数配置情况较为敏感，容易产生“过拟合”现象．

逻辑回归分类器依据特征向量到超平面决策边

界的距离来进行分类．ＬＲ 分类器首先根据已知的标注数据集进行训练，学习到一组权值 ｗ０，ｗ１，…，

ｗｍ ．当测试网页到达时，将这组权值与其特征向量

线性加和，依据式（１１）求出ｚ 值：

ｚ ＝ ｗ０ ＋ｗ１ ×ｘ１ ＋ … ＋ｗｍ ×ｘｍ （１１）

# 其中ｘ１，ｘ２，…，ｘｍ 是样本数据的各个特征，特征向量的维度为 ｍ．

然后通过ｓｉｇｍｏｄ 函数［３６］对其进行归一化，以

判别该网页属于恶意网页的概率．Ｌｅｅ 等人［３６］从

ＵＲＬ 的重定向链接关系出发，引入重定向链接长度、入口 ＵＲＬ 的出现频率等信息作为特征，在 ＬＲ 分类器的基础上提出了一种名为 ＷａｒｎｉｎｇＢｉｒｄ的近实时可疑 ＵＲＬ 监测系统，对可疑 ＵＲＬ 进行识别， 取得了９１．９０％的准确率．

逻辑回归分类器是线性分类器，因此它分类速

率快，适用于大规模数据集．此外，它还具有准确率高，可解释性好的优点．它的缺陷主要是当输入数据存在偏差的情况下分类器不收敛．

表３对比了一些主流的离线分类器的分类效果和适用范围．离线分类算法通过对整个数据集进行分析，得到全局最优的分类策略．然而，这类算法无法对数据流形式的训练集进行学习，并且处理的数据集规模始终受到内存大小的限制．

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | 表 ３ 常用离线分类器的比较 |  |
| 分类模型 | 分类速率 | 准确率 | 可解释性 数据集规模 | 局限性 |
| 朴素贝叶斯 | 高 | 低 | 好 大 | 假设特征间相互独立（往往不成立） |
| 支持向量机 | 低 | 高 | 差 小 | 分类结果对训练集的分布以及参数配置较为敏感 |
| 逻辑回归 | 高 | 高 | 好 大 | 输入数据存在偏差可能导致分类器不收敛 |

为了满足恶意网页识别的实时性要求，研究人员引入在线分类算法，每次只对一条数据进行分析， 动态调整参数以适应标注数据流的实时变化情况． 常用的在线分类算法主要包括被动攻击型算法

（ＰａｓｓｉｖｅＡｇｇｒｅｓｓｉｖｅａｌｇｏｒｉｔｈｍ，ＰＡ［４６］）、置 信度加权算法 （Ｃｏｎｆｉｄｅｎｃｅ－Ｗｅｉｇｈｔｅｄａｌｇｏｒｉｔｈｍ，ＣＷ［４７］）和权重自适应正则化算法（ＡｄａｐｔｉｖｅＲｅｇｕｌａｒｉｚａｔｉｏｎ ｏｆＷｅｉｇｈｔｓａｌｇｏｒｉｔｈｍ，ＡＲＯＷ［４８］）．

ＰＡ 算法及其变种的思想是通过引入参数 Ｔｔ和

松弛变量Ｃ 的概念来减少对分类模型中参数权值的调整．它的优点是能减少错误分类的数目，并且适用于不可分的噪声情况．而置信度加权算法 ＣＷ 认为每个学习参数都有信任度，可以用参数向量的高斯分布表示．相比于信任度大的参数，信任度小的参数更需要学习，所以会得到更频繁的修正机会．Ｍａ等人［６］收集了网页的 ＷＨＯＩＳ 信息、ＤＮＳ 信息、ＩＰ地理位置信息以及ｂｌａｃｋｌｉｓｔ信息作为特征，分别用

ＰＡ 算法和 ＣＷ 算法对可疑网页进行识别．结果显

示，ＣＷ 算法的实验效果更好，它可以将累计错误率ＣＥＲ 控制在１％ ～２％．此后，Ｃｒａｍｍｅｒ 等人［４８］对 ＣＷ 算法进行改进，提出了 ＡＲＯＷ 分类器．这种分类器的优点是抗噪性强．Ｌｅ 等人［１６］在 ＡＲＯＷ

算法的基础上，提出了用于识别钓鱼网页的系统

ＰｈｉｓｈＤｅｆ．实验证明，ＡＲＯＷ 算法在噪音１０％～３０％的情况下，累计错误率依然可以保持在１０％以内． ４．４ 基于交互式主机行为的识别方法

当访问恶意网页时，可能会出现安装恶意软件

或者执行恶意脚本的情况．这时，可以结合虚拟化技术和蜜罐技术对恶意网页进行识别．此类方法的工作原理是：使用蜜罐技术，将虚拟主机作为诱饵，访问待检测网页，通过监测访问后的主机动态行为（例如创建新进程、改变注册表及下载文件等），判断该网页是否是恶意网页．根据使用系统的不同，蜜罐技术可以细分为基于模拟的低交互式蜜罐［４９］和基于真实系统的高交互式蜜罐［５０］．诸葛建伟等人［１３］对此有详细介绍，在此不作详述．

４．５ 识别方法小结

４．１节～４．４ 节主要介绍了４ 类恶意网页识别技术，并结合已有的研究成果对其进行分析．依据上述分析结果，表４从识别方法、部署位置及评价指标等方面对这几类识别方法进行归纳总结、定性分析及横向比较．而表５ 则从识别特征及识别方法两种维度对典型的恶意网页攻击形式进行了分析和总结．

## 表 ４ 典型识别方法分类

典型工作 识别方法

部署位置

主要评价指标

黑名单 启发式规则 机器学习 主机行为 服务器端 网关 客户端 漏报率 误报率 分类速率

√ √ √ 高 低 快

ＳｍａｒｔＳｃｒｅｅｎ②

备注：ＳｍａｒｔＳｃｒｅｅｎ是基于ＩＥ 浏览器的筛选器．它根据 Ｍｉｃｒｏｓｏｆｔ提供的持续更新的 ＵＲＬ 列表及当前网页

分析到的特征，判断其是否为网络钓鱼或恶意软件，从而对用户给出提示．

√ √ 高 低 快

ＧｏｏｇｌｅＳａｆｅＢｒｏｗｓｉｎｇ③

备注：根据 Ｇｏｏｇｌｅ提供的持续更新的 ＵＲＬ 列表，允许用户检查特定 ＵＲＬ 是否存在于这个列表上，以判断

其是否为网络钓鱼或恶意软件．

√ √ √ 高 低 快

ＰｈｉｓｈＴａｎｋ

⑨

备注：ＰｈｉｓｈＴａｎｋ提供了一个自愿提交和共享钓鱼网址的开放平台，方便人们查询和识别钓鱼网页．

√ √ √ 中 低 快

ＰｈｉｓｈＮｅｔ［３８］

备注：拓宽了黑名单的识别范围，减少了漏报率．将已经发现的钓鱼 ＵＲＬ 作为先验知识，通过 ＵＲＬ 分解和

相似性计算来识别和发现新的钓鱼网页．缺点是可能引起较大的带宽消耗．

√ √ 低 高 一般

ＳｐｏｏｆＧｕａｒｄ［３９］

备注：对 ＨＴＭＬ 网页及其 ＵＲＬ（包括用户输入信息、链接关系、可疑的网址信息及图片信息等）进行检测．

可以识别一些尚未收录的恶意网页．但依赖领域知识，规则更新困难，且容易产生误判．

√ √ 低 高 一般

Ｃａｎｔｉｎａ［４１］

备注：依据网页的词频 －逆向文档频率（ＴＦ－ＩＤＦ），搜索返回结果及其他统计信息识别钓鱼网页．由于部分特

征依赖于搜索引擎的返回结果，因此会产生较大网络延迟，影响其分类速率．

√ √ 低 低 慢

Ａｕｔｏｍａｔｉｃｄｅｔｅｃｔｉｏｎｏｆ ｐｈｉｓｈｉｎｇｔａｒｇｅｔ［４２］

ＢｅｙｏｎｄＢｌａｃｋｌｉｓｔ［５］

备注：采用链接关系、关键词排序关系、文本相似性关系、层次相似性关系等作为统计特征，依据聚类算法ＤＢＳＣＡＮ 对钓鱼网页的攻击目标进行识别．这些特征的抽取需要依赖网页内容和搜索结果，因此影响其分类速率．

√ √ √ 低 低 快

备注：Ｍａ等人［５］在特征上主要抽取了 ＷＨＯＩＳ信息、ＤＮＳ信息及词汇特征．在分类方法上，分析了贝叶斯、

支持向量机及逻辑回归三类分类器在恶意网页识别问题的分类速率．

√ √ 低 低 快

ＷａｒｎｉｎｇＢｉｒｄ［３６］

备注：Ｌｅｅ等人［３６］从 ＵＲＬ 的重定向链接关系出发，提出了基于 ＬＲ 分类器的近实时恶意 ＵＲＬ 识别系统．

它的优点是分类速率快，吞吐能力强．

√ √ √ 低 低 快

ＩｄｅｎｔｉｆｙｉｎｇｓｕｓｐｉｃｉｏｕｓＵＲＬｓ［６］ 备注：Ｍａ等人采用词汇特征和主机特征，对比了多种在线分类器（例如 ＰＡ 及 ＣＷ 等）对恶意 ＵＲＬ 的识别

效果．实验证明，在累积错误率方面，ＣＷ 算法的效果优于Ｐｅｒｃｅｐｔｒｏｎ算法和ＰＡ 算法以及离线分类算法．

√ √ √ 低 低 快

ＰｈｉｓｈＤｅｆ［１６］

备注：Ｌｅ等人基于抗混淆的词汇特征和 ＡＲＯＷ 算法，提出了一种高准确率、轻量级的钓鱼网页识别系统．

实验证明，在抗噪方面，ＡＲＯＷ 算法明显优于 ＣＷ 算法．

√ √ 低 低 慢

ＨｏｓＴａＧｅ［４９］

备注：ＨｏｓＴａＧｅ是一种工作于移动端设备的低交互式的便携式蜜罐系统，主要用于检测无线网络中的恶意

软件．

√ √ √ 低 低 慢

Ｈｉｇｈｉｎｔｅｒａｃｔｉｏｎｈｏｎｅｙｐｏｔ［５０］ 备注：Ｎｉｃｏｍｅｔｔｅ等人采用高交互式的蜜罐系统，主要研究经ＳＳＨ 服务登陆宿主的入侵行为，重点分析攻击

者成功获得系统权限后的行为，填补了该方面研究的空白．

## 表 ５ 恶意网页攻击形式分类

攻击形式 钓鱼网页 恶意软件下载主机信息 √ √

ＵＲＬ 信息 √ √

识别 网页内容 √ √

特征 跳转关系 √ √

注册表变化 Ｎ／Ａ √

文件变化 Ｎ／Ａ √

黑名单 ＳｍａｒｔＳｃｒｅｅｎ② ，

② ， ③

ＧｏｏｇｌｅＳａｆｅＢｒｏｗｓｉｎｇ③ ，ＰｈｉｓｈＴａｎｋ⑨ ＳｍａｒｔＳｃｒｅｅｎ

ＧｏｏｇｌｅＳａｆｅＢｒｏｗｓｉｎｇ

识别 启发式规则 ＰｈｉｓｈＮｅｔ［３８］，ＳｐｏｏｆＧｕａｒｄ［３９］，Ｃａｎｔｉｎａ［４１］ Ｎ／Ａ

技术 Ａｕｔｏｍａｔｉｃｄｅｔｅｃｔｉｏｎｏｆｐｈｉｓｈｉｎｇｔａｒｇｅｔ［４２］，ＢｅｙｏｎｄＢｌａｃｋｌｉｓｔ［５］， ＷａｒｎｉｎｇＢｉｒｄ［３６］，ＩｄｅｎｔｉｆｙｉｎｇｓｕｓｐｉｃｉｏｕｓＵＲＬｓ［６］，ＰｈｉｓｈＤｅｆ［１６］

机器学习

ＢｅｙｏｎｄＢｌａｃｋｌｉｓｔ［５］，ＷａｒｎｉｎｇＢｉｒｄ［３６］， ＩｄｅｎｔｉｆｙｉｎｇｓｕｓｐｉｃｉｏｕｓＵＲＬｓ［６］

主机行为 Ｎ／Ａ ＨｏｓＴａＧｅ［４９］，Ｈｉｇｈｉｎｔｅｒａｃｔｉｏｎｈｏｎｅｙｐｏｔ［５０］

# 传统恶意网页识别技术面临着一些新的挑战．从来

５ 恶意网页识别面临的挑战

随着互联网的迅速发展、用户规模的不断扩大，

源上区分，这些挑战分别来自于客观环境的变化和恶意网页逃逸技术的升级．５．１ 节和５．２ 节分别介绍了这两类挑战，并就其中的技术细节进行分析和

讨论．

５．１ 客观环境变化引起的挑战

５．１．１ 网页规模大

互联网的迅速发展，使得网页规模由 ＧＢ、ＴＢ 级向ＰＢ、ＺＢ 级快速变化．巨大的网页规模对传统恶意网页识别技术提出了一些新的挑战．

首先，大量新网页的引入带来海量新特征．当使用传统机器学习算法对网页进行特征表示时，这些

新特征的引入可能产生高维特征空间，并最终导致

“维数灾难”．因此，需要设计快速有效的特征选择方法对特征进行预先筛选或引入降维方法（如主成分分析法［５１］）对高维特征空间进行处理．

其次，大规模网页限制了资源消耗大的识别方

法的应用范围．一些已有的恶意网页识别方法（例如基于主机行为的方法［１３］）需要消耗较多资源来分析主机行为或检测页面内容以判断是否为恶意网页． 这些方法的准确率和召回率高，但检测时间和资源消耗较多．随着网页规模的扩大，这些方法的局限性日益明显．当使用这些方法时，往往需要结合快速过滤器［３３］预先排除其中大多数良性网页．

最后，大规模网页对部署在网关和客户端位置

的检测工具的性能提出了更高要求．随着互联网带宽的普遍增长及网页规模的不断扩大，网关和客户端流量逐步呈现出复杂化、多样化的趋势．如何在复杂网络流量中准确识别恶意网页并及时做出响应处理，需要研究人员结合其他技术（例如高性能网包处理技术等）进行深入研究［５２］．

５．１．２ 数据集不均衡

在海量的网页中，恶意网页识别如同大海捞针，

只有极少数网页最终被确认为恶意网页．Ｇｏｏｇｌｅ的统计报告③ 显示，Ｇｏｏｇｌｅ安全浏览器每天检查数以亿计的网页，仅发现几千个不安全的站点．由此可 见，数据集的不均衡性对传统恶意网页识别方法的准确率和效率提出了严峻挑战．在特征选取层面，在 不均衡数据集产生的大量特征中，不同特征的识别效果并不相同．为了提升识别效率，研究人员设计新的识别方法，依据少量特 征快速过滤多数良性网页［５３］．在识别方法层面，不均匀的数据集启发研究人员突破传统方法的思路，寻找新的识别方法．例如，Ｉｎｖｅｒｎｉｚｚｉ等人［３３］充分利用搜索引擎查询结果，从恶意网页出发寻找相似或相近的网页，提升恶意网页所占比例，并最终提升识别效率．

５．１．３ 网页传播途径多元化

从传播途径上分析，传统的恶意网页主要依靠

电子邮件、即时通讯工具（ＩＭ）以及搜索引擎进行传播．近年来，随着社交网站的兴起和移动互联网的繁荣，通过社交网站和扫描“二维码”传播网页的行为逐渐增多．网页的传播途径因而呈现出多元化的发展趋势．这一方面拓展了恶意网页识别技术的应用领域和应用场景［５４］．另一方面，客观上对恶意网页识别技术提出了新的挑战．为此，Ｌｅｅ等人［３５－３６］从重定向关系出发提出了一个近实时检测系统，用于检测在社交网站上传播的可疑 ＵＲＬ．

在特征选取层面，这些新的应用场景的出现，催

生了更多的网络特征，丰富了特征的选择范围．在识别方法层面，多元化的网络传播途径对识别检测系统的实时性提出了更高的要求．

５．２ 恶意网页逃逸技术的升级

围绕恶意网页的攻防博弈一直在持续．本文第

４ 节主要介绍了恶意网页识别技术．为了绕过这些技术的检测，攻击者常采用环境探测＋动态加载、混淆免杀、人机识别、网页加密等技术手段来躲避检测与追踪．

５．２．１ 环境探测＋动态加载技术

随着互联网的发展，客户端浏览环境（例如浏览 器版本及插件版本等）呈现出多样性，不同浏览环境 包含的漏洞也不同．为了在提高恶意网页的攻击效率和成功率的同时保持攻击的隐蔽性，攻击者采用 了“探测页面＋攻击脚本”的“环境探测＋动态加载” 模式：即首先使用探测页面，对客户端浏览器版本、插件版本进行探测，然后决定是否使用动态加载技术（例如 ＤＯＭ ＡＰＩ等）加载攻击脚本．其中，根据攻击者资源规模的不同，可以将这种攻击细分为单攻击脚本［５５］和多攻击脚本［５６］，这里不再赘述．

５．２．２ ＵＲＬ 混淆技术

攻击者常对恶意网页的 ＵＲＬ、页面内容（攻击脚本）进行各种混淆、加密，以改变、消除其原有特征，以躲避特征扫描工具的识别．常用的混淆方式包括以下４类：

（１）在字符串中填充大量垃圾字符．

（２）改变编码方式．例如，采用十六进制编码、

Ｕｎｉｃｏｄｅ编码及ｅｓｃａｐｅ函数编码．

（３）使用ＩＰ 地址代替域名．

（４）使用随机的 ＵＲＬ 参数．

此外，恶意网页常常综合使用上述混淆方式以增强隐蔽性，提高躲避检测的成功率．

５．２．３ 人机识别技术

如３．１节介绍，在网页收集阶段，防御方常常采

用爬虫技术对网页进行主动采集．为 了躲避防御方的主动采集，攻击者常常采用一些人机识别（Ｗｅｂ

ＲｏｂｏｔＤｅｔｅｃｔｉｏｎ）技术［１２］对访问者身份进行在线判定．当认定客户端是人工浏览行为后，攻击者再执行进一步的攻击动作，否则推送事先准备好的良性网页．

其中，在线判定的方法可以分为基于图灵测试

和基于离线分析结果两类．其中基于图灵测试［５７］的方法通过在线用户答题的方式（例如要求用户输入验证码）判断用户是否是爬虫；而基于离线分析结果按照原理可分为语法日志分析、流量模式分析和分析模型训练３ 类．常见的识别依据包括：检查 ｕｓｅｒ－

ａｇｅｎｔ字段［５８］、检测有无读取ｒｏｂｏｔｓ．ｔｘｔ文件的行为［５９］等．

# 人机识别技术的使用，大大增强了恶意网页的隐蔽性，并对依赖主动采集方式的恶意网页识别方法提出了严峻的挑战．对抗此类逃逸技术的防御技术一般围绕拓宽采集方式展开．但由于采用其他采集方式受环境和资源的限制较多，因此有待深入研究．

５．２．４ 网页加密技术

一些攻击者们开始模仿正常的在线服务网站对其网页采用 ＳＳＬ 协议和 ＨＴＴＰＳ 加密服务．一 方面，采用加密服务的网页更容易取得用户信任，提高攻陷可能性；另一方面，加密恶意网页隐藏网址信息和页面内容，可以帮助逃避部署在网关的传统识别系统的检测．趋势科技的统计数据显示，２０１０ 年至 ２０１４ 年间使用 ＨＴＴＰＳ 服务的钓鱼网站瑏瑠逐年增长，从不到１０００个站点增加到超过４０００个站点．这些加密服务的使用，限制了传统依赖网址信息和页面内容的检测技术的应用范围．研究人员往往需要结合证书信息分析和检测这类恶意网页．对抗这一逃逸技术的防御技术还比较少，有待深入研究．

５．２．５ 生命周期持续缩短

部分恶意网页的生命周期持续缩短．以钓鱼网站为例，奇虎３６０ 公司的统计数据瑏瑡显示，其生存周期已经从２０１１ 年的平均５０ｈ 左右，下降到２０１２ 年下半年的不足６ｈ．恶意网页生命周期的缩短，对恶意网页识别的时效性提出新的挑战，并推动了在线识别技术的发展．

６ 研究展望

得关注和进一步探讨．

（１）针对不同应用场景的恶意网页识别方法．随着互联网的进一步发展，特别是社交网络和移动互联网［６０］的兴起和繁荣，恶意网页识别的应用场景也随之不断变化．应用场景的变化，一方面改变了恶 意网页的传播途径，另一方面对恶意网页识别的准确率和实时性［３６］提出了更高要求．为了满足这些要求，需要研究人员不断发掘识别恶意网页的特征种类，并引入新的更加适合的识别方法进行分析．因 此，这是未来恶意网页识别的可能发展方向之一．

（２）特征的比较和评测．目前用于识别恶意网

页的特征种类繁多，规模庞大．一种或一类特征可能在某一类恶意网页的识别问题上效果显著，但不适合识别其他恶意网页．为了更好地对不同特征进行比较与评测，一方面，需要构造比较合理的标注数据集；另一方面，需要在合理的特征比较和评测方法上进行研究．因此，这是未来恶意网页识别的可能发展方向之一．

（３）针对隐身逃逸技术的识别和检测．５．２ 节介绍了几种攻击者常用的隐身逃逸技术．通过使用这些技术，攻击者可以不断调整其恶意网页的外显特

征，从而规避传统方法的检测．针对特定隐身逃逸技术［６１］的识别和分类既是未来恶意网页识别的重点，也是难点．因此，这是未来恶意网页识别的可能发展 方向之一．

（４）不同识别方法的融合．不同的识别方法从不同的侧面对恶意网页进行分析和检测，选取的特

征、适用的场景各有侧重．因此，不同识别方法之间具有很强的互补性，融合各类识别方法可能是从大规模网页中识别少量恶意网页的解决途径之一．

７ 结束语

恶意网页识别是信息安全领域的热点问题．随着网络攻击技术和防御技术的不断发展，该问题一直受到研究人员的广泛关注．针对这一问题，本文首先梳理了恶意网页的基本概念，然后介绍了恶意网页识别的研究框架、应用场景和评价指标，进而对不同类别的识别方法进行深入分析和比较，分别指出它们的优势、不足以及适用场景．

上述挑战在为恶意网页识别工作带来新难题的同时，客观上也为恶意网页识别技术的新发展创造了新的条件和机遇．展望未来，仍有如下研究问题值

瑏瑠

瑏瑡

ＨＴＴＰＳ Ｐｈｉｓｈｉｎｇ ＵＲＬｓ． ｈｔｔｐ：／／ｗｗｗ．ｌｉｎｕｘｉｄｃ．ｃｏｍ／ Ｌｉｎｕｘ／２０１４－１０／１０７５５８．ｈｔｍ，２０１４．１０．２０

Ｂｒｏｗｓｅｒ Ｓｅｃｕｒｉｔｙ ａｎｄ Ｄｅｖｅｌｏｐｍｅｎｔ Ｒｅｐｏｒｔ ｏｆ Ｃｈｉｎａ． ｈｔｔｐ：／／ｚｔ．３６０．ｃｎ／ｒｅｐｏｒｔ／＃５，２０１４．１０．２０

# 本文还重点讨论了恶意网页识别面临的新挑战， 并介绍了未来恶意网页识别领域可能的研究方向．

致 谢 本文得到了国家自然科学基金委员会、国家科学技术部等机构的支持．同时，很多同行对本文的工作给予了支持和建议，在此一并表示感谢！

参 考 文 献

［１］ ＭａｈｍｏｕｄＫ，ＹｏｕｓｓｅｆＩ，Ａｎｄｒｅｗ Ｊ．Ｐｈｉｓｈｉｎｇｄｅｔｅｃｔｉｏｎ：Ａ ｌｉｔｅｒａｔｕｒｅｓｕｒｖｅｙ．ＩＥＥＥ ＣｏｍｍｕｎｉｃａｔｉｏｎｓＳｕｒｖｅｙｓ＆Ｔｕｔｏｒｉａｌｓ， ２０１３，１５（４）：２０９１－２１２１

［２］ ＰａｕｌＫ，Ｇｅｏｒｇｉａ Ｋ，ＨｅｃｔｏｒＧ Ｍ．Ｆｉｇｈｔｉｎｇｓｐａｍ ｏｎｓｏｃｉａｌ

Ｗｅｂｓｉｔｅｓ ａ ｓｕｒｖｅｙ ｏｆａｐｐｒｏａｃｈｅｓ ａｎｄ ｆｕｔｕｒｅ ｃｈａｌｌｅｎｇｅｓ． ＩＥＥＥＩｎｔｅｒｎｅｔＣｏｍｐｕｔｉｎｇ，２００７，１１（６）：３６－４５

［３］ Ｐｒｉｙａ Ｍ，Ｓａｎｄｈｙａ Ｌ，Ｃｉｚａ Ｔ．Ａ ｓｔａｔｉｃａｐｐｒｏａｃｈｔｏｄｅｔｅｃｔ

ｄｒｉｖｅ－ｂｙ－ｄｏｗｎｌｏａｄａｔｔａｃｋｓｏｎ Ｗｅｂｐａｇｅｓ／／Ｐｒｏｃｅｅｄｉｎｇｓｏｆｔｈｅ Ｉｎｔｅｒｎａｔｉｏｎａｌ Ｃｏｎｆｅｒｅｎｃｅ ｏｎ Ｃｏｎｔｒｏｌ Ｃｏｍｍｕｎｉｃａｔｉｏｎ ａｎｄ Ｃｏｍｐｕｔｉｎｇ．Ｘｉ’ａｎ，Ｃｈｉｎａ，２０１３：２９８－３０３

［４］ ＭａｖｒｏｍｍａｔｉｓＮ ＰＰ，Ｍｏｎｒｏｓｅ Ｍ Ａ Ｒ Ｆ．Ａｌｌｙｏｕｒｉｆｒａｍｅｓ

ｐｏｉｎｔｔｏ ｕｓ／／Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ｔｈｅ １７ｔｈ ＵＳＥＮＩＸ Ｓｅｃｕｒｉｔｙ Ｓｙｍｐｏｓｉｕｍ．ＳａｎＪｏｓｅ，ＵＳＡ，２００８：１－２２

［５］ ＭａＪ，ＳａｕｌＬＫ，ＳａｖａｇｅＳ，ＶｏｅｌｋｅｒＧ Ｍ．Ｂｅｙｏｎｄｂｌａｃｋｌｉｓｔｓ：

Ｌｅａｒｎｉｎｇ ｔｏ ｄｅｔｅｃｔ ｍａｌｉｃｉｏｕｓ Ｗｅｂ ｓｉｔｅｓ ｆｒｏｍ ｓｕｓｐｉｃｉｏｕｓ ＵＲＬｓ／／Ｐｒｏｃｅｅｄｉｎｇｓｏｆｔｈｅ１５ｔｈＡＣＭ ＳＩＧＫＤＤＩｎｔｅｒｎａｔｉｏｎａｌ Ｃｏｎｆｅｒｅｎｃｅｏｎ ＫｎｏｗｌｅｄｇｅＤｉｓｃｏｖｅｒｙａｎｄ Ｄａｔａ Ｍｉｎｉｎｇ．Ｎｅｗ Ｙｏｒｋ，ＵＳＡ，２００９：１２４５－１２５３

［６］ ＭａＪ，ＳａｕｌＬ Ｋ，ＳａｖａｇｅＳ，ＶｏｅｌｋｅｒＧ Ｍ．Ｉｄｅｎｔｉｆｙｉｎｇｓｕｓｐｉ－

ｃｉｏｕｓＵＲＬｓ：Ａｎａｐｐｌｉｃａｔｉｏｎｏｆｌａｒｇｅ－ｓｃａｌｅｏｎｌｉｎｅｌｅａｒｎｉｎｇ／／ Ｐｒｏｃｅｅｄｉｎｇｓｏｆｔｈｅ２６ｔｈ ＡｎｎｕａｌＩｎｔｅｒｎａｔｉｏｎａｌＣｏｎｆｅｒｅｎｃｅｏｎ ＭａｃｈｉｎｅＬｅａｒｎｉｎｇ．Ｍｏｎｔｒｅａｌ，Ｃａｎａｄａ，２００９：６８１－６８８

［７］ ＭａＪ，ＳａｕｌＬＫ，ＳａｖａｇｅＳ，ＶｏｅｌｋｅｒＧ Ｍ．Ｌｅａｒｎｉｎｇｔｏｄｅｔｅｃｔ

ｍａｌｉｃｉｏｕｓＵＲＬｓ．ＡＣＭ ＴｒａｎｓａｃｔｉｏｎｓｏｎＩｎｔｅｌｌｉｇｅｎｔＳｙｓｔｅｍｓ ａｎｄＴｅｃｈｎｏｌｏｇｙ，２０１１，２（３）：１－２４

［８］ ＣａｎａｌｉＤ，ｅｔａｌ．Ｐｒｏｐｈｉｌｅｒ：Ａｆａｓｔｆｉｌｔｅｒｆｏｒｔｈｅｌａｒｇｅ－ｓｃａｌｅ

ｄｅｔｅｃｔｉｏｎｏｆ ｍａｌｉｃｉｏｕｓ Ｗｅｂｐａｇｅｓ／／Ｐｒｏｃｅｅｄｉｎｇｓｏｆｔｈｅ２０ｔｈ ＩｎｔｅｒｎａｔｉｏｎａｌＣｏｎｆｅｒｅｎｃｅｏｎ Ｗｏｒｌｄ Ｗｉｄｅ Ｗｅｂ．Ｈｙｄｅｒａｂａｄ， Ｉｎｄｉａ，２０１１：１９７－２０６

［９］ ＴｈｏｍａｓＫ，ｅｔａｌ．Ｄｅｓｉｇｎａｎｄｅｖａｌｕａｔｉｏｎｏｆａｒｅａｌ－ｔｉｍｅＵＲＬ

ｓｐａｍｆｉｌｔｅｒｉｎｇｓｅｒｖｉｃｅ／／ＰｒｏｃｅｅｄｉｎｇｓｏｆｔｈｅＩＥＥＥ Ｓｙｍｐｏｓｉｕｍ ｏｎＳｅｃｕｒｉｔｙａｎｄＰｒｉｖａｃｙ．Ｏａｋｌａｎｄ，ＵＳＡ，２０１１：４４７－４６２

［１０］ Ｙａｄａｖ Ｓ，Ｒｅｄｄｙ Ａ Ｋ Ｋ，Ｒｅｄｄｙ Ａ Ｌ，ｅｔａｌ．Ｄｅｔｅｃｔｉｎｇ

ａｌｇｏｒｉｔｈｍｉｃａｌｌｙｇｅｎｅｒａｔｅｄ ｍａｌｉｃｉｏｕｓｄｏｍａｉｎ ｎａｍｅｓ／／Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ｔｈｅ １０ｔｈ ＡＣＭ ＳＩＧＣＯＭＭ Ｃｏｎｆｅｒｅｎｃｅ ｏｎ Ｉｎｔｅｒｎｅｔ Ｍｅａｓｕｒｅｍｅｎｔ．Ｎｅｗ Ｙｏｒｋ，ＵＳＡ，２０１０：４８－６１

［１１］ ＫｏｌｂｉｔｓｃｈＣ，ＬｉｖｓｈｉｔｓＢ，ＺｏｒｎＢ，ｅｔａｌ．Ｒｏｚｚｌｅ：Ｄｅ－ｃｌｏａｋｉｎｇ

ｉｎｔｅｒｎｅｔｍａｌｗａｒｅ／／ＰｒｏｃｅｅｄｉｎｇｓｏｆｔｈｅＩＥＥＥ Ｓｙｍｐｏｓｉｕｍ ｏｎ ＳｅｃｕｒｉｔｙａｎｄＰｒｉｖａｃｙ．ＳａｎＦｒａｎｃｉｓｃｏ，ＵＳＡ，２０１２：４４３－４５７

［１２］ Ｚｈａｎｇ Ｈｕｉ－Ｌｉｎ，Ｚｏｕ Ｗｅｉ，ＨａｎＸｉｎ－Ｈｕｉ．Ｄｒｉｖｅ－ｂｙ－ｄｏｗｎｌｏａｄ

ｍｅｃｈａｎｉｓｍｓａｎｄｄｅｆｅｎｓｅｓ．ＪｏｕｒｎａｌｏｆＳｏｆｔｗａｒｅ，２０１３，２４（４）： ８４３－８５８（ｉｎＣｈｉｎｅｓｅ）

（张慧琳，邹维，韩心慧．网页木马机理与防御技术．软件学报，２０１３，２４（４）：８４３－８５８）

［１３］ ＺｈｕｇｅＪｉａｎ－Ｗｅｉ，ＴａｎｇＹｏｎｇ，ＨａｎＸｉｎ－Ｈｕｉ，Ｄｕａｎ Ｈａｉ－Ｘｉｎ． Ｈｏｎｅｙｐｏｔｔｅｃｈｎｏｌｏｇｙ ｒｅｓｅａｒｃｈ ａｎｄ ａｐｐｌｉｃａｔｉｏｎ．Ｊｏｕｒｎａｌｏｆ Ｓｏｆｔｗａｒｅ，２０１３，２４（４）：８２５－８４２（ｉｎＣｈｉｎｅｓｅ）

（诸葛建伟，唐勇，韩心慧，段海新．蜜罐技术研究与应用进展．软件学报，２０１３，２４（４）：８２５－８４２）

［１４］ Ｘｉｏｎｇ Ｈ，ＭａｌｈｏｔｒａＰ，Ｓｔｅｆａｎ Ｄ，ｅｔａｌ．Ｕｓｅｒ－ａｓｓｉｓｔｅｄｈｏｓｔ－ ｂａｓｅｄｄｅｔｅｃｔｉｏｎｏｆｏｕｔｂｏｕｎｄ ｍａｌｗａｒｅｔｒａｆｆｉｃ／／Ｐｒｏｃｅｅｄｉｎｇｓｏｆ ｔｈｅＩｎｔｅｒｎａｔｉｏｎａｌＣｏｎｆｅｒｅｎｃｅｏｎＩｎｆｏｒｍａｔｉｏｎａｎｄＣｏｍｍｕｎｉｃａ－ ｔｉｏｎｓＳｅｃｕｒｉｔｙ．Ｂｅｉｊｉｎｇ，Ｃｈｉｎａ，２００９：２９３－３０７

［１５］ ＥｓｈｅｔｅＢ，ＶｉｌｌａｆｉｏｒｉｔａＡ，Ｗｅｌｄｅｍａｒｉａｍ Ｋ．Ｂｉｎｓｐｅｃｔ：Ｈｏｌｉｓｔｉｃ ａｎａｌｙｓｉｓａｎｄｄｅｔｅｃｔｉｏｎｏｆｍａｌｉｃｉｏｕｓ Ｗｅｂｐａｇｅｓ．ＬｅｃｔｕｒｅＮｏｔｅｓ ｏｆｔｈｅＩｎｓｔｉｔｕｔｅｆｏｒＣｏｍｐｕｔｅｒＳｃｉｅｎｃｅｓＳｏｃｉａｌＩｎｆｏｒｍａｔｉｃｓ ＆ ＴｅｌｅｃｏｍｍｕｎｉｃａｔｉｏｎｓＥｎｇｉｎｅｅｒｉｎｇ，２０１３，１０６：１４９－１６６

［１６］ Ｌｅ Ａ， Ｍａｒｋｏｐｏｕｌｏｕ Ａ， Ｆａｌｏｕｔｓｏｓ Ｍ． ＰｈｉｓｈＤｅｆ： ＵＲＬ ｎａｍｅｓｓａｙｉｔａｌｌ／／Ｐｒｏｃｅｅｄｉｎｇｓｏｆｔｈｅ３０ｔｈＩＥＥＥＩｎｔｅｒｎａｔｉｏｎａｌ ＣｏｎｆｅｒｅｎｃｅｏｎＣｏｍｐｕｔｅｒＣｏｍｍｕｎｉｃａｔｉｏｎｓ．Ｓｈａｎｇｈａｉ，Ｃｈｉｎａ， ２０１１：１９１－１９５

［１７］ ＳｈａｈｒｉａｒＨ，Ｚｕｌｋｅｒｎｉｎｅ Ｍ．Ｍｕｔｅｃ：Ｍｕｔａｔｉｏｎ－ｂａｓｅｄｔｅｓｔｉｎｇ ｏｆｃｒｏｓｓｓｉｔｅｓｃｒｉｐｔｉｎｇ／／ＰｒｏｃｅｅｄｉｎｇｓｏｆｔｈｅＩＣＳＥ Ｗｏｒｋｓｈｏｐｏｎ ＳｏｆｔｗａｒｅＥｎｇｉｎｅｅｒｉｎｇｆｏｒＳｅｃｕｒｅＳｙｓｔｅｍｓ．Ｖａｎｃｏｕｖｅｒ，Ｃａｎａｄａ， ２００９：４７－５３

［１８］ Ｗｈｉｔｔａｋｅｒ Ｃ，Ｒｙｎｅｒ Ｂ， Ｎａｚｉｆ Ｍ．Ｌａｒｇｅ－ｓｃａｌｅ ａｕｔｏｍａｔｉｃ ｃｌａｓｓｉｆｉｃａｔｉｏｎ ｏｆ ｐｈｉｓｈｉｎｇ ｐａｇｅｓ／／Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆｔｈｅ １７ｔｈ ＡｎｎｕａｌＮｅｔｗｏｒｋ ＆ ＤｉｓｔｒｉｂｕｔｅｄＳｙｓｔｅｍＳｅｃｕｒｉｔｙＳｙｍｐｏｓｉｕｍ． ＳａｎＤｉｅｇｏ，ＵＳＡ，２０１０：１－１４

［１９］ Ｒａｈｍａｎ Ｍ Ｓ，ＨｕａｎｇＴ Ｋ，ＭａｄｈｙａｓｔｈａＨ Ｖ，ｅｔａｌ．Ｅｆｆｉｃｉｅｎｔ ａｎｄｓｃａｌａｂｌｅｓｏｃｗａｒｅｄｅｔｅｃｔｉｏｎｉｎ ｏｎｌｉｎｅｓｏｃｉａｌｎｅｔｗｏｒｋｓ／／ Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ｔｈｅ ２１ｔｈ ＵＳＥＮＩＸ Ｓｅｃｕｒｉｔｙ Ｓｙｍｐｏｓｉｕｍ． Ｂｅｌｌｅｖｕｅ，ＵＳＡ，２０１２：６６３－６７８

［２０］ ＣｈｏｕＬｉ－Ｄｅｒ，ＺｈｅｎｇＨｅ，ｅｔａｌ．Ｄｅｓｉｇｎａｎｄｉｍｐｌｅｍｅｎｔａｔｉｏｎｏｆ ｃｏｎｔｅｎｔ－ｂａｓｅｄｆｉｌｔｅｒｓｙｓｔｅｍｏｎｅｍｂｅｄｄｅｄｌｉｎｕｘｈｏｍｅｇａｔｅｗａｙ

／／Ｐｒｏｃｅｅｄｉｎｇｓｏｆｔｈｅ１４ｔｈＩｎｔｅｒｎａｔｉｏｎａｌＣｏｎｆｅｒｅｎｃｅｏｎｔｈｅ

ＡｄｖａｎｃｅｄＣｏｍｍｕｎｉｃａｔｉｏｎＴｅｃｈｎｏｌｏｇｙ．Ｓｅｏｕｌ，Ｋｏｒｅａ，２０１２： １０４６－１０５１

［２１］ ＦｅｔｔｅＩ，Ｓａｄｅｈ Ｎ，Ｔｏｍａｓｉｃ Ａ．Ｌｅａｒｎｉｎｇｔｏｄｅｔｅｃｔｐｈｉｓｈｉｎｇ ｅｍａｉｌｓ／／Ｐｒｏｃｅｅｄｉｎｇｓｏｆｔｈｅ１６ｔｈＩｎｔｅｒｎａｔｉｏｎａｌＣｏｎｆｅｒｅｎｃｅｏｎ Ｗｏｒｌｄ Ｗｉｄｅ Ｗｅｂ．Ｂａｎｆｆ，Ｃａｎａｄａ，２００７：６４９－６５６

［２２］ ＳｅｉｆｅｒｔＣ，ＷｅｌｃｈＩ，ＫｏｍｉｓａｒｃｚｕｋＰ，ｅｔａｌ．Ｉｄｅｎｔｉｆｉｃａｔｉｏｎｏｆ ｍａｌｉｃｉｏｕｓＷｅｂｐａｇｅｓｔｈｒｏｕｇｈａｎａｌｙｓｉｓｏｆｕｎｄｅｒｌｙｉｎｇＤＮＳａｎｄ Ｗｅｂ ｓｅｒｖｅｒ ｒｅｌａｔｉｏｎｓｈｉｐｓ／／Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆｔｈｅ ３３ｒｄ ＩＥＥＥ ＬｏｃａｌＣｏｍｐｕｔｅｒ Ｎｅｔｗｏｒｋｓ Ｃｏｎｆｅｒｅｎｃｅ．Ｍｏｎｔｒｅａｌ，Ｃａｎａｄａ， ２００８：９３５－９４１

［２３］ Ｓｐｉｒｉｎ Ｎ， Ｈａｎ Ｊｉａ－Ｗｅｉ．Ｓｕｒｖｅｙ ｏｎ Ｗｅｂ ｓｐａｍ ｄｅｔｅｃｔｉｏｎ：

Ｐｒｉｎｃｉｐｌｅｓ ａｎｄ ａｌｇｏｒｉｔｈｍｓ． ＡＣＭ ＳＩＧＫＤＤ Ｅｘｐｌｏｒａｔｉｏｎｓ Ｎｅｗｓｌｅｔｔｅｒ，２０１２，１３（２）：５０－６４

［２４］ ＭｃＧｒａｔｈＤ Ｋ，Ｇｕｐｔａ Ｍ．Ｂｅｈｉｎｄｐｈｉｓｈｉｎｇ：Ａｎｅｘａｍｉｎａｔｉｏｎ ｏｆ ｐｈｉｓｈｅｒ ｍｏｄｉ ｏｐｅｒａｎｄｉ／／Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ｔｈｅ ＵＳＥＮＩＸ Ｗｏｒｋｓｈｏｐｏｎ Ｌａｒｇｅ－Ｓｃａｌｅ Ｅｘｐｌｏｉｔｓａｎｄ Ｅｍｅｒｇｅｎｔ Ｔｈｒｅａｔｓ． ＳａｎＪｏｓｅ，ＵＳＡ，２００８：１－８

［２５］ ＬｉａｎｇＢｉｎ，ＨｕａｎｇＪｉａｎ－Ｊｕｎ，ＬｉｕＦａｎｇ，ｅｔａｌ．ＭａｌｉｃｉｏｕｓＷｅｂ ｐａｇｅｓｄｅｔｅｃｔｉｏｎ ｂａｓｅｄ ｏｎ ａｂｎｏｒｍａｌｖｉｓｉｂｉｌｉｔｙ ｒｅｃｏｇｎｉｔｉｏｎ／／ ＰｒｏｃｅｅｄｉｎｇｓｏｆｔｈｅＩｎｔｅｒｎａｔｉｏｎａｌＣｏｎｆｅｒｅｎｃｅｏｎ Ｅ－Ｂｕｓｉｎｅｓｓ ａｎｄＩｎｆｏｒｍａｔｉｏｎＳｙｓｔｅｍＳｅｃｕｒｉｔｙ．Ｗｕｈａｎ，Ｃｈｉｎａ，２００９：１－５

［２６］ ＨａｌｌａｒａｋｅｒＯ，ＶｉｇｎａＧ．Ｄｅｔｅｃｔｉｎｇ ｍａｌｉｃｉｏｕｓＪａｖａＳｃｒｉｐｔｃｏｄｅ ｉｎ Ｍｏｚｉｌｌａ／／Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ｔｈｅ １０ｔｈ ＩＥＥＥ Ｉｎｔｅｒｎａｔｉｏｎａｌ Ｃｏｎｆｅｒｅｎｃｅｏｎ ＥｎｇｉｎｅｅｒｉｎｇｏｆＣｏｍｐｌｅｘ ＣｏｍｐｕｔｅｒＳｙｓｔｅｍｓ． Ｓｈａｎｇｈａｉ，Ｃｈｉｎａ，２００５：８５－９４

［２７］ Ｋａｌｓ Ｓ， Ｋｉｒｄａ Ｅ， Ｋｒｕｅｇｅｌ Ｃ，ｅｔ ａｌ．ＳｅｃｕＢａｔ： Ａ Ｗｅｂ ｖｕｌｎｅｒａｂｉｌｉｔｙｓｃａｎｎｅｒ／／Ｐｒｏｃｅｅｄｉｎｇｓｏｆｔｈｅ１５ｔｈＩｎｔｅｒｎａｔｉｏｎａｌ Ｃｏｎｆｅｒｅｎｃｅｏｎ Ｗｏｒｌｄ Ｗｉｄｅ Ｗｅｂ．Ｎｅｗ Ｙｏｒｋ，ＵＳＡ，２００６： ２４７－２５６

［２８］ Ｚｈｏｕ Ｌｉ， Ａｌｒｗａｉｓ Ｓ，Ｘｉｅ Ｙｉｎｇ－Ｌｉａｎ，ｅｔ ａｌ．Ｆｉｎｄｉｎｇ ｔｈｅ ｌｉｎｃｈｐｉｎｓｏｆｔｈｅｄａｒｋ Ｗｅｂ：Ａｓｔｕｄｙｏｎｔｏｐｏｌｏｇｉｃａｌｌｙｄｅｄｉｃａｔｅｄ ｈｏｓｔｓｏｎ ｍａｌｉｃｉｏｕｓ Ｗｅｂｉｎｆｒａｓｔｒｕｃｔｕｒｅｓ／／Ｐｒｏｃｅｅｄｉｎｇｓｏｆｔｈｅ ＩＥＥＥＳｙｍｐｏｓｉｕｍ ｏｎ Ｓｅｃｕｒｉｔｙａｎｄ Ｐｒｉｖａｃｙ．Ｓａｎ Ｆｒａｎｃｉｓｃｏ， ＵＳＡ，２０１３：１１２－１２６

［２９］ ＬｉＺｈｉ－Ｙｏｎｇ，Ｒａｎ Ｔａｏ，ＣａｉＺｈｅｎ－Ｈｅ，Ｚｈａｎｇ Ｈａｏ．Ａ Ｗｅｂ ｐａｇｅｍａｌｉｃｉｏｕｓｃｏｄｅｄｅｔｅｃｔａｐｐｒｏａｃｈｂａｓｅｄｏｎｓｃｒｉｐｔｅｘｅｃｕｔｉｏｎ

／／Ｐｒｏｃｅｅｄｉｎｇｓｏｆｔｈｅ５ｔｈＩｎｔｅｒｎａｔｉｏｎａｌＣｏｎｆｅｒｅｎｃｅｏｎＮａｔｕｒａｌ Ｃｏｍｐｕｔａｔｉｏｎ．Ｔｉａｎｊｉｎ，Ｃｈｉｎａ，２００９：３０８－３１２

［３０］ ＺｈａｎｇＪｕｎ－Ｊｉｅ，ＳｅｉｆｅｒｔＣ，ＳｔｏｋｅｓＪ Ｗ，Ｌｅｅ Ｗ．ＡＲＲＯＷ： Ｇｅｎｅｒａｔｉｎｇ ｓｉｇｎａｔｕｒｅｓ ｔｏ ｄｅｔｅｃｔ ｄｒｉｖｅ－ｂｙ ｄｏｗｎｌｏａｄｓ／／ Ｐｒｏｃｅｅｄｉｎｇｓｏｆｔｈｅ２０ｔｈＩｎｔｅｒｎａｔｉｏｎａｌＣｏｎｆｅｒｅｎｃｅｏｎ Ｗｏｒｌｄ Ｗｉｄｅ Ｗｅｂ．Ｈｙｄｅｒａｂａｄ，Ｉｎｄｉａ，２０１１：１８７－１９６

［３１］ ＳｅｉｆｅｒｔＣ，ＷｅｌｃｈＩ，ＫｏｍｉｓａｒｃｚｕｋＰ．Ｉｄｅｎｔｉｆｉｃａｔｉｏｎｏｆｍａｌｉｃｉｏｕｓ Ｗｅｂ ｐａｇｅｓ ｗｉｔｈ ｓｔａｔｉｃ ｈｅｕｒｉｓｔｉｃｓ／／Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ＩＥＥＥ ＣｏｎｆｅｒｅｎｃｅｏｎＴｅｌｅｃｏｍｍｕｎｉｃａｔｉｏｎＮｅｔｗｏｒｋｓａｎｄＡｐｐｌｉｃａｔｉｏｎｓ． Ａｄｅｌａｉｄｅ，Ａｕｓｔｒａｌｉａ，２００８：９１－９６

［３２］ ＷａｎｇＴａｏ，ＹｕＳｈｕｎ－Ｚｈｅｎｇ，ＸｉｅＢａｉ－Ｌｉｎ．Ａｎｏｖｅｌｆｒａｍｅｗｏｒｋ ｆｏｒｌｅａｒｎｉｎｇｔｏｄｅｔｅｃｔ ｍａｌｉｃｉｏｕｓ Ｗｅｂｐａｇｅｓ／／Ｐｒｏｃｅｅｄｉｎｇｓｏｆ ｔｈｅ Ｉｎｔｅｒｎａｔｉｏｎａｌ Ｆｏｒｕｍ ｏｎ Ｉｎｆｏｒｍａｔｉｏｎ Ｔｅｃｈｎｏｌｏｇｙ ａｎｄ Ａｐｐｌｉｃａｔｉｏｎｓ．Ｋｕｎｍｉｎｇ，Ｃｈｉｎａ，２０１０：３５３－３５７

［３３］ ＩｎｖｅｒｎｉｚｚｉＬ，ＣｏｍｐａｒｅｔｔｉＰ Ｍ，ＢｅｎｖｅｎｕｔｉＳ，ｅｔａｌ．ＥｖｉｌＳｅｅｄ： ＡｇｕｉｄｅｄａｐｐｒｏａｃｈｔｏｆｉｎｄｉｎｇｍａｌｉｃｉｏｕｓＷｅｂｐａｇｅｓ／／Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆｔｈｅ ＩＥＥＥ Ｓｙｍｐｏｓｉｕｍ ｏｎ Ｓｅｃｕｒｉｔｙ ａｎｄ Ｐｒｉｖａｃｙ． Ｓａｎ Ｆｒａｎｃｉｓｃｏ，ＵＳＡ，２０１２：４２８－４４２

［３４］ ＢａｔｉｓｔａＧＥ，ＰｒａｔｉＲＣ，ＭｏｎａｒｄＭ Ｃ．Ａｓｔｕｄｙｏｆｔｈｅｂｅｈａｖｉｏｒ ｏｆｓｅｖｅｒａｌ ｍｅｔｈｏｄｓｆｏｒｂａｌａｎｃｉｎｇ ｍａｃｈｉｎｅｌｅａｒｎｉｎｇｔｒａｉｎｉｎｇ ｄａｔａ．ＡＣＭ ＳＩＧＫＤＤＥｘｐｌｏｒａｔｉｏｎｓＮｅｗｓｌｅｔｔｅｒ，２００４，６（１）： ２０－２９

［３５］ ＬｅｅＳ，Ｋｉｍ Ｊ．ＷａｒｎｉｎｇＢｉｒｄ：Ｄｅｔｅｃｔｉｎｇｓｕｓｐｉｃｉｏｕｓ ＵＲＬｓｉｎ ｔｗｉｔｔｅｒｓｔｒｅａｍ／／Ｐｒｏｃｅｅｄｉｎｇｓｏｆｔｈｅ１９ｔｈ ＡｎｎｕａｌＮｅｔｗｏｒｋ ＆ ＤｉｓｔｒｉｂｕｔｅｄＳｙｓｔｅｍ ＳｅｃｕｒｉｔｙＳｙｍｐｏｓｉｕｍ．Ｓａｎ Ｄｉｅｇｏ，ＵＳＡ， ２０１２：１－１３

［３６］ ＬｅｅＳ， Ｋｉｍ Ｊ． ＷａｒｎｉｎｇＢｉｒｄ： Ａ ｎｅａｒ ｒｅａｌ－ｔｉｍｅ ｄｅｔｅｃｔｉｏｎ ｓｙｓｔｅｍ ｆｏｒ ｓｕｓｐｉｃｉｏｕｓ ＵＲＬｓ ｉｎ ｔｗｉｔｔｅｒ ｓｔｒｅａｍ． ＩＥＥＥ Ｔｒａｎｓａｃｔｉｏｎｓｏｎ Ｄｅｐｅｎｄａｂｌｅａｎｄ Ｓｅｃｕｒｅ Ｃｏｍｐｕｔｉｎｇ，２０１３， １０（３）：１８３－１９５

［３７］ Ｈｏｎｅｙｎｅｔ Ｐｒｏｊｅｃｔ． Ｋｎｏｗ Ｙｏｕｒ Ｅｎｅｍｙ： Ｌｅａｒｎｉｎｇ ａｂｏｕｔ Ｓｅｃｕｒｉｔｙ Ｔｈｒｅａｔｓ．２ｎｄ Ｅｄｉｔｉｏｎ． Ｂｏｓｔｏｎ： Ａｄｄｉｓｏｎ－Ｗｅｓｌｅｙ Ｐｒｏｆｅｓｓｉｏｎａｌ，２００４

［３８］ ＰｒａｋａｓｈＰ，Ｋｕｍａｒ Ｍ，Ｋｏｍｐｅｌｌａ Ｒ Ｒ，ｅｔａｌ．ＰｈｉｓｈＮｅｔ： Ｐｒｅｄｉｃｔｉｖｅｂｌａｃｋｌｉｓｔｉｎｇｔｏｄｅｔｅｃｔｐｈｉｓｈｉｎｇａｔｔａｃｋｓ／／Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆｔｈｅ ２９ｔｈ ＩＥＥＥ Ｉｎｔｅｒｎａｔｉｏｎａｌ Ｃｏｎｆｅｒｅｎｃｅ ｏｎ Ｃｏｍｐｕｔｅｒ Ｃｏｍｍｕｎｉｃａｔｉｏｎｓ．ＳａｎＤｉｅｇｏ，ＵＳＡ，２０１０：１－５

［３９］ ＳｈｅｎｇＳ，ＷａｒｄｍａｎＢ，ＷａｒｎｅｒＧ，ｅｔａｌ．Ａｎｅｍｐｉｒｉｃａｌａｎａｌｙｓｉｓ ｏｆｐｈｉｓｈｉｎｇｂｌａｃｋｌｉｓｔｓ／／Ｐｒｏｃｅｅｄｉｎｇｓｏｆｔｈｅ６ｔｈＣｏｎｆｅｒｅｎｃｅｉｎ ＥｍａｉｌａｎｄＡｎｔｉ－Ｓｐａｍ．Ｍｏｕｎｔａｉｎｖｉｅｗ，ＵＳＡ，２００９：１－１０

［４０］ ＣｈｏｕＮ，ＬｅｄｅｓｍａＲ，ＴｅｒａｇｕｃｈｉＹ，ＭｉｔｃｈｅｌｌＪＣ．Ｃｌｉｅｎｔ－ｓｉｄｅ ｄｅｆｅｎｓｅ ａｇａｉｎｓｔ Ｗｅｂ－ｂａｓｅｄ ｉｄｅｎｔｉｔｙ ｔｈｅｆｔ／／Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ｔｈｅ１１ｔｈ Ａｎｎｕａｌ Ｎｅｔｗｏｒｋ ＆ Ｄｉｓｔｒｉｂｕｔｅｄ Ｓｙｓｔｅｍ Ｓｅｃｕｒｉｔｙ Ｓｙｍｐｏｓｉｕｍ．ＳａｎＤｉｅｇｏ，ＵＳＡ，２００４：１－１６

［４１］ Ｚｈａｎｇ Ｙｕｅ，ＨｏｎｇＪＩ，Ｃｒａｎｏｒ Ｌ Ｆ．Ｃａｎｔｉｎａ：Ａ ｃｏｎｔｅｎｔ－ ｂａｓｅｄａｐｐｒｏａｃｈｔｏｄｅｔｅｃｔｉｎｇｐｈｉｓｈｉｎｇ Ｗｅｂｓｉｔｅｓ／／Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆｔｈｅ１６ｔｈＩｎｔｅｒｎａｔｉｏｎａｌＣｏｎｆｅｒｅｎｃｅｏｎ Ｗｏｒｌｄ Ｗｉｄｅ Ｗｅｂ． Ｂａｎｆｆ，Ｃａｎａｄａ，２００７：６３９－６４８

［４２］ ＬｉｕＧａｎｇ，Ｑｉｕ Ｂｉｔｅ，Ｌｉｕ Ｗｅｎ－Ｙｉｎ．Ａｕｔｏｍａｔｉｃｄｅｔｅｃｔｉｏｎｏｆ ｐｈｉｓｈｉｎｇｔａｒｇｅｔｆｒｏｍ ｐｈｉｓｈｉｎｇ Ｗｅｂｐａｇｅ／／Ｐｒｏｃｅｅｄｉｎｇｓｏｆｔｈｅ ２０ｔｈ Ｉｎｔｅｒｎａｔｉｏｎａｌ Ｃｏｎｆｅｒｅｎｃｅ ｏｎ Ｐａｔｔｅｒｎ Ｒｅｃｏｇｎｉｔｉｏｎ． Ｉｓｔａｎｂｕｌ，Ｔｕｒｋｅｙ，２０１０：４１５３－４１５６

［４３］ ＥｓｔｅｒＭ，ＫｒｉｅｇｅｌＨ－Ｐ ，ＳａｎｄｅｒＪ，ＸｕＸｉａｏ－Ｗｅｉ．Ａｄｅｎｓｉｔｙ－ ｂａｓｅｄａｌｇｏｒｉｔｈｍｆｏｒｄｉｓｃｏｖｅｒｉｎｇｃｌｕｓｔｅｒｓｉｎｌａｒｇｅｓｐａｔｉａｌｄａｔａ－ ｂａｓｅｓ ｗｉｔｈ ｎｏｉｓｅ／／Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆｔｈｅ ２ｎｄ ＡＣＭ ＳＩＧＫＤＤ ＩｎｔｅｒｎａｔｉｏｎａｌＣｏｎｆｅｒｅｎｃｅｏｎ Ｋｎｏｗｌｅｄｇｅ Ｄｉｓｃｏｖｅｒｙａｎｄ Ｄａｔａ Ｍｉｎｉｎｇ．Ｐｏｒｔｌａｎｄ，ＵＳＡ，１９９６：２２６－２３１

［４４］ Ｈｕａｎｇ Ｈ，Ｑｉａｎ Ｌ， Ｗａｎｇ Ｙ．Ａ ＳＶＭ－ｂａｓｅｄｔｅｃｈｎｉｑｕｅｔｏ ｄｅｔｅｃｔ ｐｈｉｓｈｉｎｇ ＵＲＬｓ．Ｉｎｆｏｒｍａｔｉｏｎ Ｔｅｃｈｎｏｌｏｇｙ Ｊｏｕｒｎａｌ， ２０１２，１１（７）：９２１－９２５

［４５］ ＫｏｌａｒｉＰ，ＦｉｎｉｎＴ，ＪｏｓｈｉＡ．ＳＶＭｓｆｏｒｔｈｅＢｌｏｇｏｓｐｈｅｒｅ：Ｂｌｏｇ ｉｄｅｎｔｉｆｉｃａｔｉｏｎａｎｄｓｐｌｏｇｄｅｔｅｃｔｉｏｎ／／Ｐｒｏｃｅｅｄｉｎｇｓｏｆｔｈｅ ＡＡＡＩ ＳｐｒｉｎｇＳｙｍｐｏｓｉｕｍｏｎＣｏｍｐｕｔａｔｉｏｎａｌＡｐｐｒｏａｃｈｅｓｔｏＡｎａｌｙｚｉｎｇ Ｗｅｂｌｏｇｓ．ＭｅｎｌｏＰａｒｋ，ＵＳＡ，２００６：９２－９９

［４６］ Ｃｒａｍｍｅｒ Ｋ，Ｄｅｋｅｌ Ｏ，ＫｅｓｈｅｔＪ，ｅｔａｌ．Ｏｎｌｉｎｅ ｐａｓｓｉｖｅ－ ａｇｇｒｅｓｓｉｖｅａｌｇｏｒｉｔｈｍｓ．ＪｏｕｒｎａｌｏｆＭａｃｈｉｎｅＬｅａｒｎｉｎｇＲｅｓｅａｒｃｈ， ２００６，７：５５１－５８５

［４７］ Ｂｌｕｍ Ａ，ＷａｒｄｍａｎＢ，ＳｏｌｏｒｉｏＴ，ＷａｒｎｅｒＧ．Ｌｅｘｉｃａｌｆｅａｔｕｒｅ ｂａｓｅｄ ｐｈｉｓｈｉｎｇ ＵＲＬ ｄｅｔｅｃｔｉｏｎ ｕｓｉｎｇ ｏｎｌｉｎｅ ｌｅａｒｎｉｎｇ／／ Ｐｒｏｃｅｅｄｉｎｇｓｏｆｔｈｅ３ｒｄ ＡＣＭ Ｗｏｒｋｓｈｏｐｏｎ ＡｒｔｉｆｉｃｉａｌＩｎｔｅｌｌｉ－ ｇｅｎｃｅａｎｄＳｅｃｕｒｉｔｙ．Ｃｈｉｃａｇｏ，ＵＳＡ，２０１０：５４－６０

［４８］ ＣｒａｍｍｅｒＫ，ＫｕｌｅｓｚａＡ，Ｄｒｅｄｚｅ Ｍ．Ａｄａｐｔｉｖｅｒｅｇｕｌａｒｉｚａｔｉｏｎ ｏｆｗｅｉｇｈｔｖｅｃｔｏｒｓ．ＭａｃｈｉｎｅＬｅａｒｎｉｎｇ，２０１３，９１（２）：１５５－１８７

［４９］ ＶａｓｉｌｏｍａｎｏｌａｋｉｓＥ，ＫａｒｕｐｐａｙａｈＳ，Ｆｉｓｃｈｅｒ Ｍ，ｅｔａｌ．Ｔｈｉｓ ｎｅｔｗｏｒｋｉｓｉｎｆｅｃｔｅｄ：ＨｏｓＴａＧｅ－ａｌｏｗ－ｉｎｔｅｒａｃｔｉｏｎｈｏｎｅｙｐｏｔｆｏｒ ｍｏｂｉｌｅｄｅｖｉｃｅｓ／／Ｐｒｏｃｅｅｄｉｎｇｓｏｆｔｈｅ３ｒｄ ＡＣＭ Ｗｏｒｋｓｈｏｐｏｎ ＳｅｃｕｒｉｔｙａｎｄＰｒｉｖａｃｙｉｎＳｍａｒｔｐｈｏｎｅｓ ＆ ＭｏｂｉｌｅＤｅｖｉｃｅｓ．Ｓａｎ Ｆｒａｎｃｉｓｃｏ，ＵＳＡ，２０１３：４３－４８

［５０］ Ｎｉｃｏｍｅｔｔｅ Ｖ， Ｋａｎｉｃｈｅ Ｍ， Ａｌａｔａ Ｅ，ｅｔ ａｌ．Ｓｅｔ－ｕｐ ａｎｄ ｄｅｐｌｏｙｍｅｎｔｏｆａｈｉｇｈ－ｉｎｔｅｒａｃｔｉｏｎｈｏｎｅｙｐｏｔ：Ｅｘｐｅｒｉｍｅｎｔａｎｄ ｌｅｓｓｏｎｓｌｅａｒｎｅｄ．ＪｏｕｒｎａｌｉｎＣｏｍｐｕｔｅｒＶｉｒｏｌｏｇｙ，２０１１，７（２）： １４３－１５７

［５１］ Ｂｉｓｈｏｐ Ｃ Ｍ．Ｐａｔｔｅｒｎ Ｒｅｃｏｇｎｉｔｉｏｎ ａｎｄ Ｍａｃｈｉｎｅ Ｌｅａｒｎｉｎｇ． Ｎｅｗ Ｙｏｒｋ：Ｓｐｒｉｎｇｅｒ，２００６

［５２］ Ｑｉ Ｙａ－Ｘｕａｎ，ＬｉＪｕｎ． Ｔｈｅｏｒｅｔｉｃａｌａｎａｌｙｓｉｓ ａｎｄ ａｌｇｏｒｉｔｈｍ ｄｅｓｉｇｎｏｆｈｉｇｈ－ｐｅｒｆｏｒｍａｎｃｅｐａｃｋｅｔｃｌａｓｓｉｆｉｃａｔｉｏｎａｌｇｏｒｉｔｈｍｓ． ＣｈｉｎｅｓｅＪｏｕｒｎａｌｏｆ Ｃｏｍｐｕｔｅｒｓ，２０１３，３６（２）：４０８－４２１（ｉｎ Ｃｈｉｎｅｓｅ）

（亓亚烜，李军．高性能网包分类理论与算法综述．计算机学报，２０１３，３６（２）：４０８－４２１）

［５３］ Ｌｉｎ Ｍｉｎ－Ｓｈｅｎｇ，Ｃｈｉｕ Ｃｈｉｅｎ－Ｙｉ，Ｌｅｅ Ｙｕｈ－Ｊｙｅ，Ｐａｏ Ｈｓｉｎｇ－

Ｋｕｏ． Ｍａｌｉｃｉｏｕｓ ＵＲＬ ｆｉｌｔｅｒｉｎｇ—Ａ ｂｉｇ ｄａｔａ ａｐｐｌｉｃａｔｉｏｎ／／ ＰｒｏｃｅｅｄｉｎｇｓｏｆｔｈｅＩＥＥＥ Ｉｎｔｅｒｎａｔｉｏｎａｌ Ｃｏｎｆｅｒｅｎｃｅ ｏｎ Ｂｉｇ Ｄａｔａ．ＳａｎｔａＣｌａｒａ，ＵＳＡ，２０１３：５８９－５９６

［５４］ Ｗｈｌｉｓｃｈ Ｍ，ＴｒａｐｐＳ，ＫｅｉｌＣ，ｅｔａｌ．Ｆｉｒｓｔｉｎｓｉｇｈｔｓｆｒｏｍ ａ

ｍｏｂｉｌｅ ｈｏｎｅｙｐｏｔ／／Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ｔｈｅ ＡＣＭ ＳＩＧＣＯＭＭ Ｃｏｎｆｅｒｅｎｃｅ ｏｎ Ａｐｐｌｉｃａｔｉｏｎｓ，Ｔｅｃｈｎｏｌｏｇｉｅｓ， Ａｒｃｈｉｔｅｃｔｕｒｅｓ， ａｎｄ Ｐｒｏｔｏｃｏｌｓ ｆｏｒ Ｃｏｍｐｕｔｅｒ Ｃｏｍｍｕｎｉｃａｔｉｏｎ． Ｈｅｌｓｉｎｋｉ， Ｆｉｎｌａｎｄ，２０１２：３０５－３０６

［５５］ ＫａｐｒａｖｅｌｏｓＡ，ＳｈｏｓｈｉｔａｉｓｈｖｉｌｉＹ，Ｃｏｖａ Ｍ，ｅｔａｌ．Ｒｅｖｏｌｖｅｒ：

Ａｎａｕｔｏｍａｔｅｄ ａｐｐｒｏａｃｈ ｔｏｔｈｅ ｄｅｔｅｃｔｉｏｎ ｏｆｅｖａｓｉｖｅ Ｗｅｂ－ ｂａｓｅｄｍａｌｗａｒｅ／／Ｐｒｏｃｅｅｄｉｎｇｓｏｆｔｈｅ２２ｎｄ ＵＳＥＮＩＸ Ｓｅｃｕｒｉｔｙ Ｓｙｍｐｏｓｉｕｍ．Ｗａｓｈｉｎｇｔｏｎ，ＵＳＡ，２０１３：６３７－６５２

［５６］ Ｃｈｅｎ Ｋｅｖｉｎ Ｚｈｉｊｉｅ，Ｇｕ Ｇｕｏ－Ｆｅｉ，Ｚｈｕｇｅ Ｊｉａｎ－Ｗｅｉ，ｅｔ ａｌ．

ＷｅｂＰａｔｒｏｌ：Ａｕｔｏｍａｔｅｄｃｏｌｌｅｃｔｉｏｎａｎｄｒｅｐｌａｙ ｏｆ Ｗｅｂ－ｂａｓｅｄ ｍａｌｗａｒｅｓｃｅｎａｒｉｏｓ／／Ｐｒｏｃｅｅｄｉｎｇｓｏｆｔｈｅ６ｔｈＡＣＭ Ｓｙｍｐｏｓｉｕｍ ｏｎＩｎｆｏｒｍａｔｉｏｎ， Ｃｏｍｐｕｔｅｒａｎｄ Ｃｏｍｍｕｎｉｃａｔｉｏｎｓ Ｓｅｃｕｒｉｔｙ．

Ｓｉｎｇａｐｏｒｅ，２０１１：１８６－１９５

［５７］ Ｄｏｒａｎ Ｄ，Ｇｏｋｈａｌｅ Ｓ Ｓ． Ｗｅｂ ｒｏｂｏｔｄｅｔｅｃｔｉｏｎ ｔｅｃｈｎｉｑｕｅｓ： Ｏｖｅｒｖｉｅｗ ａｎｄ ｌｉｍｉｔａｔｉｏｎｓ． Ｄａｔａ Ｍｉｎｉｎｇ ａｎｄ Ｋｎｏｗｌｅｄｇｅ Ｄｉｓｃｏｖｅｒｙ，２０１１，２２（１）：１８３－２１０

［５８］ ＳｔｅｖａｎｏｖｉｃＤ，Ｖｌａｊｉｃ Ｎ，Ａｎ Ａ．Ｄｅｔｅｃｔｉｏｎｏｆ ｍａｌｉｃｉｏｕｓａｎｄ

ｎｏｎ－ｍａｌｉｃｉｏｕｓ Ｗｅｂｓｉｔｅ ｖｉｓｉｔｏｒｓ ｕｓｉｎｇ ｕｎｓｕｐｅｒｖｉｓｅｄ ｎｅｕｒａｌ ｎｅｔｗｏｒｋｌｅａｒｎｉｎｇ．Ａｐｐｌｉｅｄ Ｓｏｆｔ Ｃｏｍｐｕｔｉｎｇ，２０１３，１３（１）： ６９８－７０８

［５９］ ＬｅｅＪＹ，ＬｅｅＹ Ｔ．Ａｆｒａｍｅｗｏｒｋｆｏｒａｒｅｓｅａｒｃｈｉｎｖｅｎｔｏｒｙｏｆ

ｓｕｓｔａｉｎａｂｉｌｉｔｙａｓｓｅｓｓｍｅｎｔｉｎ ｍａｎｕｆａｃｔｕｒｉｎｇ．ＪｏｕｒｎａｌｏｆＣｌｅａｎｅｒ Ｐｒｏｄｕｃｔｉｏｎ，２０１４，７９：２０７－２１８

［６０］ Ｙａｎｇ Ｈｕａｎ，Ｚｈａｎｇ Ｙｕ－Ｑｉｎｇ，Ｈｕ Ｙｕ－Ｐｕ，Ｌｉｕ Ｑｉ－Ｘｕ．Ａ

ｍａｌｗａｒｅｂｅｈａｖｉｏｒｄｅｔｅｃｔｉｏｎｓｙｓｔｅｍ ｏｆ Ａｎｄｒｏｉｄａｐｐｌｉｃａｔｉｏｎｓ ｂａｓｅｄｏｎｍｕｌｔｉ－ｃｌａｓｓｆｅａｔｕｒｅｓ．ＣｈｉｎｅｓｅＪｏｕｒｎａｌｏｆＣｏｍｐｕｔｅｒｓ， ２０１４，３７（１）：１５－２７（ｉｎＣｈｉｎｅｓｅ）

（杨欢，张玉清，胡予濮，刘奇旭．基于多类特征的 Ａｎｄｒｏｉｄ

应用恶意行为检测系统．计算机学报，２０１４，３７（１）：１５－２７）

［６１］ Ｓｈａ Ｈｏｎｇ－Ｚｈｏｕ，Ｚｈｏｕ Ｚｈｏｕ，Ｌｉｕ Ｑｉｎｇ－Ｙｕｎ， Ｑｉｎ Ｐｅｎｇ． Ｌｉｇｈｔ－ｗｅｉｇｈｔｓｅｌｆ－ｌｅａｒｎｉｎｇｆｏｒＵＲＬｃｌａｓｓｉｆｉｃａｔｉｏｎ．Ｊｏｕｒｎａｌｏｎ Ｃｏｍｍｕｎｉｃａｔｉｏｎｓ，２０１４，３５（９）：３２－３９（ｉｎＣｈｉｎｅｓｅ）

（沙泓州，周舟，刘庆云，秦鹏．轻量级的自学习网页分类方

法．通信学报，２０１４，３５（９）：３２－３９）

## ＳＨＡ Ｈｏｎｇ－Ｚｈｏｕ，ｂｏｒｎ ｉｎ １９８８， Ｐｈ．Ｄ．ｃａｎｄｉｄａｔｅ．Ｈｉｓｒｅｓｅａｒｃｈｉｎｔｅｒｅｓｔｓ ｉｎｃｌｕｄｅ ｉｎｆｏｒｍａｔｉｏｎ ｓｅｃｕｒｉｔｙ， ｎｅｔｗｏｒｋ ｓｅｃｕｒｉｔｙ．

ＬＩＵ Ｑｉｎｇ－Ｙｕｎ，ｂｏｒｎｉｎ１９８０，Ｐｈ．Ｄ．，ｓｅｎｉｏｒｅｎｇｉｎｅｅｒ． Ｈｉｓｒｅｓｅａｒｃｈｉｎｔｅｒｅｓｔｓｉｎｃｌｕｄｅｉｎｆｏｒｍａｔｉｏｎｓｅｃｕｒｉｔｙ，ｎｅｔｗｏｒｋ ｓｅｃｕｒｉｔｙ．

ＬＩＵＴｉｎｇ－Ｗｅｎ，ｂｏｒｎｉｎ１９８６，Ｐｈ．Ｄ．，ａｓｓｉｓｔａｎｔｒｅｓｅａｒｃｈｅｒ．

Ｈｉｓｒｅｓｅａｒｃｈｉｎｔｅｒｅｓｔｓｉｎｃｌｕｄｅ ｂｉｇ ｄａｔａ ｓｅｃｕｒｉｔｙ ａｎａｌｙｓｉｓ，

ａｌｇｏｒｉｔｈｍｄｅｓｉｇｎａｎｄａｎａｌｙｓｉｓ．

ＺＨＯＵ Ｚｈｏｕ，ｂｏｒｎｉｎ１９８３，Ｐｈ．Ｄ．，ｓｅｎｉｏｒｅｎｇｉｎｅｅｒ． Ｈｉｓ ｒｅｓｅａｒｃｈ ｉｎｔｅｒｅｓｔｓ ｉｎｃｌｕｄｅ ｎｅｔｗｏｒｋ ｓｅｃｕｒｉｔｙ， ｈｉｇｈ－ ｐｅｒｆｏｒｍａｎｃｅｎｅｔｗｏｒｋ．

ＧＵＯ Ｌｉ，ｂｏｒｎｉｎ１９６９，ｓｅｎｉｏｒｅｎｇｉｎｅｅｒ．Ｈｅｒｒｅｓｅａｒｃｈ

ｉｎｔｅｒｅｓｔｓｉｎｃｌｕｄｅｉｎｆｏｒｍａｔｉｏｎｓｅｃｕｒｉｔｙ，ｎｅｔｗｏｒｋｓｅｃｕｒｉｔｙａｎｄ ｄａｔａｓｔｒｅａｍｐｒｏｃｅｓｓｉｎｇ．

ＦＡＮＧＢｉｎ－Ｘｉｎｇ，ｂｏｒｎｉｎ１９６０，Ｐｈ．Ｄ．，ｐｒｏｆｅｓｓｏｒ，Ｐｈ．Ｄ．

ｓｕｐｅｒｖｉｓｏｒ，ｍｅｍｂｅｒｏｆ Ｃｈｉｎｅｓｅ Ａｃａｄｅｍｙ ｏｆ Ｅｎｇｉｎｅｅｒｉｎｇ． Ｈｉｓｃｕｒｒｅｎｔｒｅｓｅａｒｃｈｉｎｔｅｒｅｓｔｓｉｎｃｌｕｄｅｎｅｔｗｏｒｋｓｅｃｕｒｉｔｙａｎｄ ｉｎｆｏｒｍａｔｉｏｎｃｏｎｔｅｎｔｓｅｃｕｒｉｔｙ．

Ｂａｃｋｇｒｏｕｎｄ

Ｗｉｔｈｔｈｅｒａｐｉｄｄｅｖｅｌｏｐｍｅｎｔｏｆｎｅｔｗｏｒｋｔｅｃｈｎｏｌｏｇｙ，ｔｈｅ ｍａｌｉｃｉｏｕｓｗｅｂｐａｇｅｄｅｔｅｃｔｉｏｎｈａｓｂｅｃｏｍｅｍｕｃｈｍｏｒｅｉｍｐｏｒｔａｎｔ ｔｈａｎｅｖｅｒｉｎｔｈｅｆｉｅｌｄｏｆｎｅｔｗｏｒｋｓｅｃｕｒｉｔｙ．Ｄｕｒｉｎｇｒｅｃｅｎｔｙｅａｒｓ， ａｇｒｅａｔｎｕｍｂｅｒｏｆｒｅｓｅａｒｃｈｅｒｓｈａｖｅｐａｉｄ ｍｕｃｈａｔｔｅｎｔｉｏｎｔｏ ｔｈｉｓａｒｅａ，ａｎｄ ｆｏｃｕｓｅｄ ｏｎ ｔｈｅｏｒｅｔｉｃａｌ ｍｏｄｅｌａｎｄ ｐｒａｃｔｉｃａｌ ａｐｐｌｉｃａｔｉｏｎｓ．Ｈｏｗｅｖｅｒ，ａｌｏｔｏｆｉｓｓｕｅｓｓｔｉｌｌｈａｖｅｎｏｔｂｅｅｎ ａｄｄｒｅｓｓｅｄｗｅｌｌ．

Ｔｈｕｓ，ｔｈｉｓｐａｐｅｒｆｉｒｓｔｓｔｕｄｉｅｓｔｈｅｂａｓｉｃｃｏｎｃｅｐｔｏｆｍａｌｉ－

ｃｉｏｕｓｗｅｂｐａｇｅｓａｎｄｔｈｅｎｉｎｔｒｏｄｕｃｅｓｔｈｅｒｅｓｅａｒｃｈｆｒａｍｅｗｏｒｋ， ａｐｐｌｉｃａｔｉｏｎｓｃｅｎａｒｉｏｓａｎｄ ｅｖａｌｕａｔｉｏｎｉｎｄｅｘ ｏｆｔｈｅ ｄｅｔｅｃｔｉｏｎ ｐｒｏｃｅｓｓ．Ｆｕｒｔｈｅｒｍｏｒｅ，ｉｔ ｍａｋｅｓｉｎ－ｄｅｐｔｈ ａｎａｌｙｓｉｓ ａｍｏｎｇ

ｄｉｆｆｅｒｅｎｔｃａｔｅｇｏｒｉｅｓｏｆｄｅｔｅｃｔｉｏｎ ａｐｐｒｏａｃｈｅｓａｎｄ ｐｏｉｎｔｓｏｕｔ ｔｈｅｉｒ ａｄｖａｎｔａｇｅｓ，ｄｉｓａｄｖａｎｔａｇｅｓ ｓｅｐａｒａｔｅｌｙ． Ｂｅｓｉｄｅｓ，ｔｈｉｓ ｐａｐｅｒａｌｓｏ ｄｉｓｃｕｓｓｅｓｔｈｅ ｎｅｗ ｃｈａｌｌｅｎｇｅｓｉｎｔｈｉｓａｒｅａ，ａｎｄ ｉｎｔｒｏｄｕｃｅｓｔｈｅｐｏｓｓｉｂｌｅｒｅｓｅａｒｃｈｄｉｒｅｃｔｉｏｎｉｎｔｈｉｓｆｉｅｌｄｉｎｔｈｅ ｆｕｔｕｒｅ．

Ｔｈｉｓ ｗｏｒｋ ｉｓ ｓｕｐｐｏｒｔｅｄ ｂｙ ｔｈｅ Ｓｔｒａｔｅｇｉｃ Ｐｒｉｏｒｉｔｙ Ｒｅｓｅａｒｃｈ Ｐｒｏｇｒａｍ ｏｆ ｔｈｅ Ｃｈｉｎｅｓｅ Ａｃａｄｅｍｙ ｏｆ Ｓｃｉｅｎｃｅｓ

（Ｎｏ．ＸＤＡ０６０３０２００），ｔｈｅ ＮａｔｉｏｎａｌＳｃｉｅｎｃｅａｎｄ Ｔｅｃｈｎｏｌｏｇｙ

Ｓｕｐｐｏｒｔ Ｐｒｏｇｒａｍ （Ｎｏ．２０１２ＢＡＨ４６Ｂ０２）ａｎｄ ｔｈｅ Ｎａｔｉｏｎａｌ ＮａｔｕｒａｌＳｃｉｅｎｃｅＦｏｕｎｄａｔｉｏｎｏｆＣｈｉｎａ （Ｎｏ．６１４０２４７４）．