多因素综合评定在移动终端网络钓鱼检测中的应用

徐欢潇[1] 祁娟[2] 陈枢茜[1] 林佩园[1] 苏萍[1]

（1. 南通理工学院计算机与信息工程学院 江苏 南通 226019；2 南通理工学院 传媒与设计学院 江苏 南通 226019）

摘要：在深入研究大量反钓鱼技术文献和现存的钓鱼网站识别机制的基础上，针对移动终端的特点，本文在多因素综合评分模块结合了传统的黑白名单过滤模块的实验结果、URL 检测模块的实验结果以及页面特征检测模块的实验结果，其综合评定结果进一步说明了本文提出的方法的优越性。

关键字：移动终端，网络钓鱼，多因素综合评定

一、引言

据中国互联网网络信息中心（CNNIC）的第 48 次报告[1]，截至 2021 年 6 月，我国网民使用手机上网的比例达 99.6%，这就使得针对手机用户的钓鱼攻击量持续增长，利用钓鱼手段在手机端进行欺诈的活动呈爆发趋势。图 1-1 是来自 CNNIC 的互联网络接入设备使用情况。国际反网络钓鱼工作组官网 APWG 的报告指出[2]，APWG 在 2021 年 6 月看到了 222127 次袭击，这是 APWG 报告历史上最坏的一个月。而通过对现有的反钓鱼系统识别流程的研究，发现目前存在的网络钓鱼攻击识别技术使用的分类算法单一以及评价偏重正确率的问题。为了降低钓鱼识别检测的误判率和漏报率，本文根据移动终端的应用特点，提出一种新的，面向移动应用的，多因素综合评定的反钓鱼检测方法。

图 1-1 互联网络接入设备使用情况

二、基于混合的检测

随着钓鱼网站越来越复杂，人们越来越难以用单一的方式去处理它，因此越来越多的研究人员开始尝试从多技术结合的角度去考虑问题。典型的有 Jayshree Hajgud 等结合了黑/白名单和启发式这三种技术来减少误报率[3]。

基于混合的检测实质上就是将已有的好的方法进行有机结合，选取最适合的组合方法来提高检测的效果。鉴于各种方法的使用局限性，多技术混合技术成为今后的发展趋势。为了找到最好的组合方式，需要充分了解现有技术，了解用于钓鱼网站识别的特征及识别原理、性能和适应性，进行多信息的融合、多技术并发的方案设计。

三、多因素综合评价模块设计

本文提出的一种网络钓鱼识别系统，跟传统的检测方法基本一致[4]，与其他方法相比，最重要的也是最不一样的部分是在页面特征识别检测模块，本文采用了一种多特征整合、多技术并发，在线测评和多级测评兼顾的钓鱼检测引擎。此外，在多因素综合评分模块结合了本文提出的黑白名单过滤模块的实验结果、URL 检测模块的实验结果以及页面特征检测模块的实验结果，其综合评分进一步说明了本文提出的方法的优越性。

1、模块设计

多因素综合评分模块的评分主要参照黑/白名单过滤模块、URL 特征检测模块和页面特征检测模块的评分结果，在此基础上进行综合性的评分。

本模块的设计主要分别两个阶段，首先分别对黑/白名单过滤模块、URL 特征检测模块和页面特征检测模块进行模块设计，选取合适的属性值的集合，然后根据各特征值的不同得到所在模块的检测结果。最后，根据三大模块的检测结果的综合性分析研究，选取合适的特征值集合用于多因素综合模块的评分，得到最后的检测结果。

（1）黑/白名单过滤模块

本文使用 $V<U,S>$ 表示一条 URL 的属性，$U<u_{black},u_{white}>$ 代表该 URL 各属性特征值的集合，各特征值的计算如表 1-1 所示，$S<s_{similarity}>$ 代表该 URL 的检测结果。

表 1-1 URL 过滤中各特征值的计算

$U.u_{black}$	0	URL 在黑名单库中
	1	URL 不在黑名单库中
$U.u_{white}$	0	URL 在白名单库中
	1	URL 不在白名单库中
$S.s'_{similarity}$	0	该条 URL 经过分类结果判定为可信网站
	1	该条 URL 经过分类结果判定为钓鱼网站

（2）URL 特征评分模块

本文使用 $V'<U',S'>$ 表示一条 URL 的签名，$U'<u'_{ip},u'_{special},u'_{domain},u'_{path},u'_{whois}>$ 代表该 URL 各属性特征值的集合，各特征值的计算如表 1-2 所示，$S'<s'_{similarity}>$ 代表该 URL 的检测结果。

表 1-2 URL 签名中各特征值的计算

$U'.u'_{ip}$	0	URL 的全域名不含 IP 形式

	1	URL 的全域名含有 IP 形式
$U'.u'_{special}$	0	URL 中不含特殊字符
	1	URL 中含有特殊字符
$U'.u'_{domain}$	0	URL 的域名级数 < 5
	1	URL 的域名级数≥5
$U'.u'_{path}$	0	URL 的路径数 < 5
	1	URL 的路径数≥5
$U'.u'_{whois}$	0	URL 的 whois 在 2021 年之前注册
	1	URL 的 whois 在 2021 年注册
$S'.s'_{similarity}$	0	该条 URL 经过分类结果判定为可信网站
	1	该条 URL 经过分类结果判定为钓鱼网站

（3）页面特征评分模块

本文使用 $V'' < U'', S'' >$ 表示 URL 的属性，$U'' < u''_T, u''_V, u''_{TP} >$ 代表各特征属性特征值的集合，各特征值的计算如表 1–3 所示，$S'' < s''_{similarity} >$ 代表该 URL 的检测结果。

表 1–3 页面各特征对应特征值的计算

$U''.u''_T$	0	文本特征的相似度相对较高
	1	文本特征的相似度相对较低
$U''.u''_V$	0	视觉特征的相似度相对较高
	1	视觉特征的相似度相对较低
$U''.u''_{TP}$	0	拓扑特征的相似度相对较高
	1	拓扑特征的相似度相对较低
$S''.s''_{simila}$	0	该条 URL 经过分类结果判定为可信网站
	1	该条 URL 经过分类结果判定为钓鱼网站

（4）综合评分模块

综合黑/白名单模块、URL 特征评分模块和页面特征评分模的评分结果，本文使用 $V''' < U''', S''' >$ 表示 URL 的属性，$U''' < u'''_L, u'''_U, u'''_P >$ 代表各类特征属性特征值的集合，各类特征值的计算如表 1–4 所示，$S''' < s'''_{similarity} >$ 代表该 URL 的检测结果。

表 1–4 页面各类特征对应特征值的计算

$U'''.u'''_L$	0	非黑 or 非白
	1	黑 or 白
$U'''.u'''_U$	0	没有超过预定的阈值
	1	超过了预定的阈值
$U'''.u'''_P$	0	没有被模板库中命中
	1	被模板库中命中
$S'''.s'''_{similarity}$	0	该条 URL 经过分类结果判定为可信网站
	1	该条 URL 经过分类结果判定为钓鱼网站

2、模块实现

综合评分模块的评分标准如表 1–5 所示，根据不同的黑/白名单特征（$U'''.u'''_L$）、URL 特征（$U'''.u'''_U$）和页面特征（$U'''.u'''_P$）的值，最后得到的结果分为 8 种。

表 1–5 综合评分准则

0	0	0	Non–phish
0	1	0	用户指引
0	0	1	Phish
1	0	0	Non–phish/Phish
1	0	1	Non–phish/Phish
1	1	0	Non–phish/Phish
1	1	1	Non–phish/Phish

四、多因素综合评价识别检测结果

本文提出的多特征分类融合的方法在一定程度上比常规方法存在显著优势，但是为了能进一步强化这一优点。本文再一次采用融合的思想，把前面涉及的包括黑/白名单过滤模块的结果、URL 特征检测模块的结果和页面特征检测模块的结果都融入到检测结果的判断条件里。此处针对这一想法提出了融合以上所有模块检测结果的多因素综合评分模块。当然，越多的因素被考虑进来越好，但是考虑到实际工作量以及预测分析等原因，本文仅选用本文所提方法的因素进行考虑，该模型下测试数据集的实验结果如表 1–6 所示：

样本总数 W	总体识别率 ACC	正向样本检出率 P–ACC	总误报率 FR	总漏报率 MR
2000	99.10%	99.03%	1.20%	0.10%
2000	98.40%	97.60%	1.70%	0.30%
2000	98.10%	97.60%	1.90%	0.50%
2000	98.60%	92.10%	2.00%	1.70%

表 1–6 基于多因素综合评价算法实验结果

显而易见，本文采用的基于多因素综合评分算法进一步优化了钓鱼网页的识别率，跟常规的仅使用基于页面特征识别检测算法相比较，其在漏报率方面相对来说差不多，但是在正确率方面提高很多，识别效率显著提升。特别地，其在误报率方面更是效果显著，这是对于单因素（页面特征识别检测模块实验结果）评分模块的非常好的改进与完善，之后可以考虑更多的因素。

参考文献：

[1]中国互联网络信息中心(CNNIC).中国互联网络状况发展统计报告[R/OL].[2021–09].

[2]Anti–Phishing Working Group (APWG). Phishing Activity Trends Report, 2nd Quarter 2021 [R/OL].

[3]Jayshree Hajgude,Lata Ragha. “Phish Mail Guard :Phishing Mail Detection Technique by using Textual and URL Analysis”.[C] World Congress on Information and Communication Technologies(WICT),2012, pp: 297–302

[4]吴朝花. 基于 Android 平台的网络钓鱼识别系统的设计与实现[D]. 北京: 北京邮电大学, 2012.

依托项目：南通市科技（指导性）项目“多源异构数据融合分析技术在移动网络钓鱼攻击检测中的应用研究”（编号：JCZ20141）