一种基于集成学习的钓鱼网站检测方法

余恩泽,努尔布力,于 清
新疆大学 信息科学与工程学院,乌鲁木齐 830046

摘 要:针对钓鱼攻击者常用的伪造HTTPS网站以及其他混淆技术,借鉴了目前主流基于机器学习以及规则匹配的检测钓鱼网站的方法RMLR和PhishDef,增加对网页文本关键字和网页子链接等信息进行特征提取的过程,提出了Nmap-RF分类方法。Nmap-RF是基于规则匹配和随机森林方法的集成钓鱼网站检测方法。根据网页协议对网站进行预过滤,若判定其为钓鱼网站则省略后续特征提取步骤。否则以文本关键字置信度,网页子链接置信度,钓鱼类词汇相似度以及网页PageRank作为关键特征,以常见URL、Whois、DNS信息和网页标签信息作为辅助特征,经过随机森林分类模型判断后给出最终的分类结果。实验证明,Nmap-RF集成方法可以在平均9~10 μs的时间内对钓鱼网页进行检测,且可以过滤掉98.4%的不合法页面,平均总精度可达99.6%。

关键词:钓鱼网页;集成学习;规则匹配;钓鱼网页混淆技术
文献标志码:A **中图分类号:**TP393.08 doi:10.3778/j.issn.1002-8331.1812-0362

余恩泽,努尔布力,于清.一种基于集成学习的钓鱼网站检测方法.计算机工程与应用,2019,55(18):81-88.
YU Enze, Nurbol, YU Qing. Phishing website detection method based on integrated learning. Computer Engineering and Applications, 2019, 55(18):81-88.

Phishing Website Detection Method Based on Integrated Learning

YU Enze, Nurbol, YU Qing
College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China

Abstract: In view of the fake HTTPS websites commonly used by phishing attackers and other obfuscation techniques, this paper draws on the current mainstream methods of detecting phishing websites based on machine learning and rule matching, RMLR and PhishDef, and adds features such as web page text keywords and web page sub-links. The Nmap-RF classification method is proposed. Nmap-RF is an integrated phishing website detection method based on rule matching and random forest method. The website is pre-filtered according to the webpage protocol, and if it is determined to be a phishing website, the subsequent feature extraction step is omitted. Otherwise, the text keyword confidence, the page sub-link confidence, the phishing vocabulary similarity and the page PageRank are taken as key features. The common URL, Whois, DNS information and web page tag information are used as auxiliary features, and are judged by the random forest classification model. Experiments show that the Nmap-RF integration method can detect phishing pages in an average of 9~10 μs, and can filter out 98.4% of illegal pages. The average total accuracy is 99.6%.

Key words: phishing websites; ensemble learning; rule matching; phishing obfuscation techniques

1 引言

钓鱼(phishing)攻击是指以仿冒银行、电子商务平台或其他知名机构身份,发送以引诱收信人个人敏感信息(如真实姓名,相关账户口令、数字ID,ATM PIN码或其他个人数据)为目的虚假网页,从而获取经济或其他形式利益的一种网络攻击形式。在现实生活中,钓鱼攻击造成的经济损失是巨大的,且近年来呈增长之势。文献[1-2]指出仅在2018年上半年,360互联网安全中心共

基金项目:国家自然科学基金(No.61433012,No.61562082,No.61303231)。
作者简介:余恩泽(1999—),男,本科生,研究领域为网络安全,机器学习,自然语言处理,E-mail:yez_aisec@163.com;努尔布力(1981—),男,博士,教授,研究领域为网络安全,数据挖掘;于清(1973—),女,副教授,研究领域为自然语言处理,机器翻译,人工智能。
收稿日期:2018-12-28 **修回日期:**2019-03-20 **文章编号:**1002-8331(2019)18-0081-08
CNKI**网络出版:**2019-05-31, http://kns.cnki.net/kcms/detail/11.2127.tp.20190530.0922.004.html

截获各类新增钓鱼网站1 622.6万个,同比2017年上半年(201.5万个)上升7倍;平均每天新增9.0万个。文献[3]指出仅在2018年8月份一个月内,中国反钓鱼网站联盟共处理钓鱼网站2 400个。随着万维网的快速发展,目前常见的钓鱼攻击主要是攻击者利用人们的心理弱点,诱骗用户进入自己精心构建的网站,从而获得用户输入的隐私信息。鉴于钓鱼网页造成的巨大损失,如何应对日益增长的网络钓鱼技术已经成为网络安全研究的热点话题之一。

钓鱼网站检测主要从网页的URL字符串,网页文本内容以及DNS信息,Whois信息和关键字进行特征提取从而进行检测。随着技术的发展,攻击者也相应有了规避检测的方式。如URL混淆技术,网站虚假注册信息,页面包含大量合法网页超链接,集中注册重复的DNS,虚假的域名IP所有者信息,对URL进行混淆处理等。同时,传统意义上安全的HTTPS协议网站也并不完全安全,因为通过对数字证书进行自签名的方式以及构造虚假证书等方法,攻击者可以迷惑用户,规避钓鱼网站检测工具,以达到攻击目的。因此,针对钓鱼混淆技术展开研究具有重要意义。

本文对从PhishTank和Yahoo两个网站获取到的两类URL数据集进行分析,数据集共包含10 975个网站。与现有研究方法较少考虑到HTTPS和HTTP的区别不同,本文针对HTTPS提出了三条预过滤规则,再结合子链接信息以及文本内容等特征提出Nmap-RF方法。

2 相关工作

2.1 基于黑白名单的钓鱼网站检测方法

此类检测方法是一种简单有效的钓鱼网站检测方法。网站黑名单和网站白名单中分别保存了已被确认为钓鱼网站或合法网站的域名,URL等信息。当用户对某一网站进行访问时,该网页的URL或域名就会自动在黑白名单中进行匹配,根据匹配结果做出相应允许访问和禁止访问的判断。这种方法很少产生误报,且实现简单。然而它往往需要人工进行复查,而且这种方法过度依赖于黑白名单的规模,误报率高,无法检查处理新出现的钓鱼网页。文献[4]显示,约有93%的钓鱼网页没有被主流的黑名单收录,如何存储和定期更新发布有效的黑白名单成为了关键的问题。

2.2 基于搜索引擎的钓鱼网站检测方法

传统的基于关键字相似度匹配的钓鱼网站检测方法是基于搜索引擎来实现的。通过使用TF-IDF算法提取网页文本关键字,依据网页文本关键字在搜索因此的搜索结果链接的合法性判断该网页的性质。如文献[5]提出的CANTINA检测方法,该方法在提取了 n 个关键词后进行预判断,如果搜索结果中靠前的内容里出现了原URL或者域名,则判定该网页为合法网页。这种方法的缺陷很明显,攻击者可以在网页文本中加入大量无关且在合法网页中出现频率高的词汇进行躲避检测。而且搜索引擎对结果中网页的排名具有不确定性,这往往会导致误判。

2.3 基于机器学习的钓鱼网站检测方法

基于机器学习的钓鱼网站检测方法是把该问题当成了样本二分类问题或者聚类问题,此类方法的流程一般为,首先对样本进行特征提取,如URL字符信息,DNS信息。文献[6]还对网页图像信息进行特征提取等。其次再用样本选择不同算法训练出分类器,最后使用分类器对网页进行判定。文献[7]通过对网页Logo进行图像处理,进而判定网页性质。Li等人[8]使用了球形SVM算法检测钓鱼网页,Hu等人[9]利用五个分类器和四个集成学习分类器检测网页合法性,Martínez等人[10]通过比较,以条件互信息做度量。再使用聚类算法对网站做出分类。与传统的两种检测方法相比,该类检测方法可以实现对新的钓鱼网站进行检测,具有较好的泛化能力。但是泛化能力的好坏依赖于特征的选择,这往往需要专家进行决策,而且此类方法如果提取的向量过多,检测时时间复杂度就会比其他方法高。

2.4 基于集成方法的钓鱼网站检测方法

这一类检测方法参考了机器学习中集成学习的概念,集成学习的思想,主要是针对不同类型的特征,构造了不同的分类学习模型。集成学习中最终分类结果可以通过加权投票法确定。文献[11]和文献[12]通过使用集成学习的方法,提取不同的特征信息并以此训练出不同的基础分类器模型,最后利用分类集成策略综合多个基础分类器生成最终的结果,准确率最高可达97.2%。

3 Nmap-RF集成方法

图1所示为Nmap-RF的检测框架,具体的检测步骤如下:

步骤1 判断网页是否使用HTTPS协议,若未使用则进入步骤4,否则进入步骤2。

步骤2 使用Nmap对待检网页的URL进行证书信息收集。

步骤3 证书信息是否匹配规则中任意一条规则,若匹配则将网页标签记为钓鱼网页并进入步骤6,若未匹配则进入步骤4。

步骤4 提取待检URL的URL特征、DNS特征、Whois特征、PageRank特征、文本特征、源代码特征、子链接特征并构成特征向量。

步骤5 将待检URL的特征向量输入训练好的RF分类器中,得出分类标签。

步骤6 根据分类标签得到网页性质,结束。

3.1 基于数字证书的预过滤

针对使用HTTPS协议的网站,传统的检测方法先

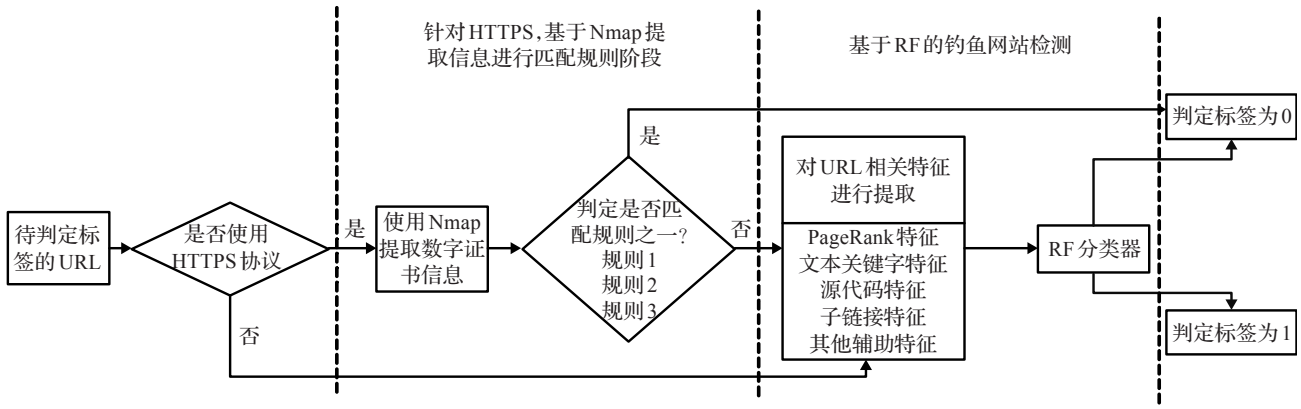


图1 Nmap-RF检测框架

对数字证书信息进行特征提取,再使用机器学习分类方法进行分类。这种方法往往时延较高,针对此现象,本文先使用Nmap对网站证书信息进行收集,再对证书信息进行规则匹配。Nmap是一款网络连接端扫描软件,常用来扫描某个主机或者网络开放的端口。进而推断运行的服务。根据Zheng等人[13]和Clark等人[14]对数字证书信息特征提取的研究,结合数据集在相关特征的取值,本文提出了三条预过滤规则。

规则1 数字证书版本不是V3直接判定该网站为非法钓鱼网站。

数字证书的版本表明了证书的标准结构,X.509V3是目前证书的最新版本,合法的HTTPS网站都会选择此版本的证书。非法的HTTPS网站由于成本原因,X.509V1和X.509V2,根据数字认证中心要求,V1版本的数字证书已经不推荐使用,而V2版本由于自身的局限性同样没有得到推广应用,因此本文提出了规则1去检测此类混淆手段。

规则2 采用RSA加密算法的数字证书,若其密钥长度低于1 024位,则直接判定该网站为非法网站。

RSA加密算法是目前数字证书最常用的加密方法之一,而在当今的计算条件下,密钥长度小于1 024位的密钥都被认为是不安全的。因此本文提出了规则2去检测此类混淆手段。

规则3 证书截止时间和注册时间间距低于2个月的网站直接判定为非法。

恶意的数字证书中,其证书截止时间和注册时间往往会很短,而正常的HTTPS网页一般都在一年以上。因此本文提出了规则3去应对这一类混淆手段。

实验结果表明,这三条预处理规则帮助过滤了14.6%的恶意HTTPS网站,能够基于数字证书的部分特征做有效预过滤,且没有误报。这省去了部分不完善的非法网页需要进行特征提取的时间,进一步减少检测钓鱼网站模型所需的时间,满足了对于实时性检测的需求。

3.2 基于网页基础信息的特征提取

3.2.1 基于网页URL的特征提取

URL(Uniform Resource Locator),即统一资源定位符。URL是可以唯一地标识互联网中某个资源的网络地址字符串。一般的URL可以分成五个不同的字段:即协议、域名、路径、文件名以及参数。钓鱼攻击者往往会在其中插入大量不相关词汇混淆用户的视觉。针对这一现象,Ding等人[15]通过提取URL字符串长度,各类字符的信息熵作为网页的检测特征。本文参考文献[16]也以信息熵指标以及其他URL特征作为URL特征提取的一个依据。

信息熵是发生一个事件时所得到的信息量的平均,其大小可用来描述某个信息的不确定度。针对随机生成垃圾字符的钓鱼URL字符串,本文提取URL中的字母和数字的信息熵作为网页检测特征。信息熵计算公式如式(1)所示:

$$H=-\sum_{i=0}^{n}p_{x_i}\mathrm{lb}\,p_{x_i} \tag{1}$$

其中 p_{x_i} 表示某一事件发生的概率。其他特征从URL词汇部分、域名部分、路径部分、文件部分和参数部分进行特征提取。

表1所示为一个URL的详细的特征,综上,可以针对一条URL的字符串特征定义为 V_u:

$$V_u=<f_1,f_2,\cdots,f_{38}> \tag{2}$$

其中 f_n 代表第 n 个特征。

3.2.2 基于网页PageRank的特征提取

PageRank又被称作网页级别,是通过一个网页中众多的超链接的关系来确定一个网页的等级,PageRank是Google等搜索引擎对网页进行排名的重要依据。其取值一般在0~9之间,0表示网页的重要性最低,9表示重要度最高。一般可以通过PageRank取值判断一个网页的受欢迎程度,例如Google会把自己定为9。因此本文以网页PageRank作为一个待提取特征 V_p,且只使用源URL进行一次提取PageRank查询,忽略网页文本中出现的URL。该特征可定义为:

$$V_p=\begin{cases}\text{网页}PageRank,(PageRank\text{可查询})\\-1,(PageRank\text{不可查询})\end{cases} \tag{3}$$

表1 URL特征

类别	特征项	数量
URL词汇部分	最长词汇长度,字母的数目,词汇个数,最长词汇占URL总长的比例,最长词汇与URL平均词汇长差值,纯字母词汇个数,纯数字词汇个数,大小写混用字母词汇个数,词汇包含大写字母的数量,字母和数字混用词汇在总词汇数所占比例,字母占URL长度的比例,数字占URL长度的比例,dot的数目,非数字和字母的个数,是否包括奇异字符,包含奇异字符的数目,包含%的字符数目,%字符占URL总长度的比例	18
URL域名部分	域名+协议的原始长度,域名长度,含有dot的数目,词汇数量,是否有端口号,单个词汇的最大长度,连接符的数目,路径部分长度和文件名长度的差值	8
URL路径部分	路径中连接符的数目,路径中的dot的数目,路径中最长词汇的长度	3
URL文件部分	文件名的长度,文件中连接符的数目,文件中含有dot的个数	3
URL参数部分	参数字符串的长度,参数字符串连接符的数目,参数部分含词汇的数目,参数部分最长词汇的长度最长词汇长度、参数部分连接符数量	4

3.2.3 基于网页Whois信息的特征提取

Whois(Who is)是现在互联网中不可或缺的一项信息服务,它用来查询某一域名的IP一级其所有者的信息。Whois信息用来描述一个网页的注册信息:网页的注册人信息、注册地点、注册时间、失效时间。

由于钓鱼网站注册具有成本低,并发性强等特点,往往攻击者在注册钓鱼网站时会使用相同或类似的信息进行注册。据此本文提出了基于网页Whois信息的特征提取方法。

Whois信息特征集合定义如下:注册时间的年份,月份以及日期,Whois信息更新时间的年份、月份以及日期。过期时间的年份、月份以及日期。注册的IP是否存在于IP黑名单中。登记者注册信息是否存在于已知的黑名单中。是否是私人注册。共计11个特征。

在实际的特征提取中,约有四分之一的网站Whois信息无法查询。为此,Ma等人[17]指出,包括Whois特征的特征集合在多种分类器里的准确性最高,即使排除了Whois特征,也可以出现高度准确的分类结果。通过4.4.3小节的实验结果,本文仍然提取URL的Whois特征。待检测URL的Whois特征向量定义为V_W。

$$V_W=<f_1,f_2,\cdots,f_{11}> \tag{4}$$

其中f_n代表第n个特征。

3.2.4 基于网页DNS信息的特征提取

DNS(Domain Name System)是保存域名和IP映射关系的服务器,它可描述网页提供相关解析服务的信息。如该域名下所有的IP地址的记录,别名记录以及邮件路由记录等。Le等人[18]通过分析了DNS和Web服务器与网页间的关系检测恶意网页,取得了良好的效果。Zhang等人[19]通过引入时间窗口的概念,检测网页在一段时间内实际使用的DNS服务器更新相关DNS服务器的DNS质疑度来对网站进行检测。据此,与文献[16]不同的是,增加了NS_NUM和MX_NUM记录,并对无法解析的DNS信息记录增加了平滑处理措施,实验证明,增加了平滑处理后,准确率和召回率各有1.4%和1.2%的上升。

DNS信息特征集合定义如下:IP个数、DNS置信度、RETRY值和集合最大RETRY值的比值,REFRESH值和集合最大REFRESH值的比值,EXPIRE值和集合最大EXPIRE值的比值,TTL值和集合最大TTL值的比值,NS_NUM值和集合最大NU_NUM值的比值,IP是否存在于黑名单列表中。共计八个特征。

以NS_NUM为例,平滑处理的公式如下:

$$f_{\text{ns\_num}}=\begin{cases} ns\_num & \text{if}(ns\_num!=0) \\ \dfrac{1}{ns\_num} & \text{if}(ns\_num==0) \end{cases} \tag{5}$$

其中,ns_num表示网站集合中不同取值的类别个数。待检测URL的DNS特征向量定义为V_D。

$$V_D=<f_{\text{ip\_num}},\cdots,f_{0\text{-}1}> \tag{6}$$

其中$f_{\text{ip_num}}$代表IP数目,$f_{0\text{-}1}$代表是否存在于黑名单中。

3.3 基于网页源代码的特征提取

3.3.1 基于网页标签信息的特征提取

超文本标记语言标记标签是组成网页的基本元素,攻击者往往为了达到攻击效果,在编写网站时会较多地使用无意义的死链接,修改网页标题等方法去迷惑其他用户。基于文献[16]的标签提取工作,本文提取了<Meta>,<Title>,<a>,<link>等标签作为特征,并在其中加入拉普拉斯平滑处理,以减少未在样本出现过的标签词汇对实验结果的影响。

本文提取网页HTML的17个特征,并将待检URL的HTML特征向量定义为V_H。

$$V_H=<f_{H_{\text{meta\_1}}},f_{H_{\text{meta\_2}}},\cdots,f_{\text{link\_domain\_dot}}> \tag{7}$$

3.3.2 基于网页子链接信息的特征提取

网页链接是网页开发者在编写网页时对内外链接网站的引用。它分为内链和外链,内链是指指向本站内部的链接,一般会在同一个域名下,而外链是指外部网站指向本站的链接。在社交网站数据挖掘中,某个用户经常转发或者点赞的网页链接往往是这个用户所关心的,它们最能代表这个用户的喜好。与之类似,在钓鱼

网站给出的子链接中,往往会链接到其他的恶意页面中,据此本文提出了基于网页子链接信息的特征提取方法。

网页经常使用的链接标签有两种,第一种是以<a href=>标签为主的链接形式,另一种是以<link rel=>标签为主的链接形式。例如贴吧和<link rel="search" type="application/opensearchdescription+xml" href="/content-search.xml" title="百度搜索"/>就是这样的例子。

本文通过提取出合法网站集合和钓鱼网站集合中所有rel和href属性后的网站作为集合,提出了下列四种特征。

将某一网站子链接在合法网页子链接域名集合出现的次数定义为 $f_{\text{n\_num}}$:

$$f_{\text{n\_num}}=\begin{cases}n\_num & \text{if}(w \text{ in } S\_domain)\\ 0 & \text{if}(w \text{ not in } S\_domain)\end{cases} \tag{8}$$

将某一网站子链接在非法网页子链接域名集合中出现的次数定义为 $f_{\text{e\_num}}$:

$$f_{\text{e\_num}}=\begin{cases}e\_num & \text{if}(w \text{ in } S\_e\_domain)\\ 0 & \text{if}(w \text{ not in } S\_e\_domain)\end{cases} \tag{9}$$

将某一网站子链接在合法网页子链接域名集合中出现的频率定义为 $f_{\text{p\_n\_num}}$:

$$f_{\text{p\_n\_num}}=\begin{cases}\dfrac{n\_num}{|S\_domain|} & \text{if}(w \text{ in } S\_domain)\\ 0 & \text{if}(w \text{ not in } S\_domain)\end{cases} \tag{10}$$

将某一网站子链接在非法网页子链接域名集合中出现的频率定义为 $f_{\text{p\_e\_num}}$:

$$f_{\text{p\_e\_num}}=\begin{cases}\dfrac{e\_num}{|S\_e\_domain|} & \text{if}(w \text{ in } S\_e\_domain)\\ 0 & \text{if}(w \text{ not in } S\_e\_domain)\end{cases} \tag{11}$$

其中,n_num 表示某一网站子链接在合法网页子链接域名集合中出现的次数,e_num 表示某一网站子链接在非法网页子链接域名集合出现的次数,S_domain 表示合法网页子链接域名集合,S_e_domain 表示非法网页子链接域名集合。将待检测URL的子链接特征向量定义为:

$$V_{\text{son}}=<f_{\text{n\_num}},f_{\text{e\_num}},f_{\text{p\_n\_num}},f_{\text{p\_e\_num}}> \tag{12}$$

3.3.3 基于文本关键词信息的特征提取

现存的钓鱼网站检测方法中,对于网页文本内容信息进行特征提取的研究较少。一方面是因为内容信息文本庞杂,冗余度和噪声较高。另一方面是因为网页文本内容往往较难提取分词。根据本文基于TF-IDF加权技术首先对钓鱼网页文本内容出现频率高的词汇进行提取,再计算每个网页中文本信息在该词汇集合中出现的频率,将其作为特征,具体的流程图如图2。

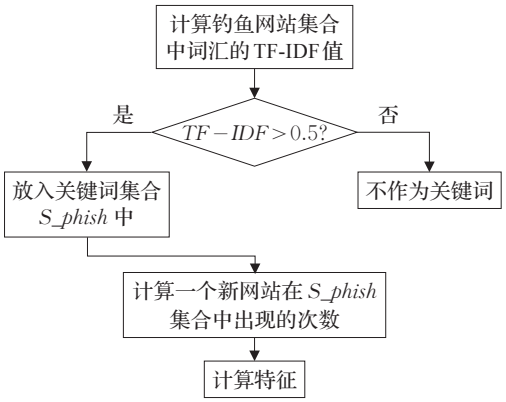


图2 文本信息内容特征提取

TF-IDF(Term Frequency-Inverse Document Frequency)是一种常用的文本数据挖掘的加权技术。它的基本思想是,如果某一个单词在该文本内容中出现的频率很高,并且在其他文本内容中出现的频率较低,则可以认为该单词具有较好的区分这两篇文章的能力。这样的思想可以有效地帮助应用到进行分类问题中,本文根据实验,使用0.5作为关键阈值频率对词汇进行区分,TF-IDF的具体计算公式如下:

$$TF(X_i,j)=\frac{Count(j,X_i)}{Count(j)} \tag{13}$$

$$IDF(X_i)=\lg\left(\frac{N}{Files\_count(X_i)}\right) \tag{14}$$

通过TF-IDF筛选出区分钓鱼网站和合法网站的关键词,获得了一个包含394个关键词集合,将它记为 S_{phish}。再通过计算某个网页文本集合出现在这个集合中的次数得到 $f_{\text{p\_text}}$,其计算公式如下:

$$f_{\text{p\_text}}=Count(p\_text) \tag{15}$$

将待检测URL的子链接特征向量定义为:

$$V_s=<f_r> \tag{16}$$

3.3.4 基于网页文本类钓鱼词汇的特征提取

钓鱼攻击者以设备硬件电气信号、系统时间、系统信息或其他类别信息作为随机序列的种子,并按规律随机产生URL字符串。虽然该类随机生成URL字符串的方法攻击成本相对较低,但它却具有良好的攻击效果。基于[16]选择两类词汇作为钓鱼词汇库词汇,即域名标签以及形如"数字字符-字母字符-数字字符"或"字母字符-数字字符-字母字符"的词汇纳入钓鱼类词汇库再进行比较的工作,本文比较了Jaccard距离、编辑距离、莱文斯坦比、Jaro距离以及Jaro_winkler距离这5种计算字符串相似度的方法,最终选择了速率最快的莱文斯坦比。

Jaccard相似度计算公式为:

$$JM=\frac{|str1\cap str2|}{|str1\cup str2|} \tag{17}$$

其中‖符号表示集合中元素的个数,$str1$ 和 $str2$ 是两个需要比较的字符串。

编辑距离没有具体的计算公式,它的大小等于由 *str*1 转化成 *str*2 最少的操作次数,可使用的操作包括插入、删除、替换。

莱文斯坦比的计算公式为:

$$r=\frac{sum-ldist}{sum} \tag{18}$$

其中 *sum* 是指 *str*1 和 *str*2 字串的长度总和,*ldist* 是类编辑距离。

Jaro距离的的计算公式为:

$$d_j=\begin{cases}0 & \text{if}(m==0)\\ \frac{1}{3}\left(\frac{m}{|str1|}+\frac{m}{|str2|}+\frac{m-t}{m}\right) & \text{otherwise}\end{cases} \tag{19}$$

其中 m 为 *str*1 和 *str*2 的匹配长度。当两个字符串在同一个位置的字符相等则认为匹配,t 为换位次数,它等于不同顺序的匹配字符的数目的一半。

Jaro_winkler距离的计算公式为:

$$d_w=d_j+(lp(1-d_j)) \tag{20}$$

其中 p 是调整分数的常数,d_j 是两个字符串的Jaro距离,l 是前缀的相同的长度,规定最大为4。

本文没有使用阈值直接判定网页性质,而是将待检测网页的URL与钓鱼类词汇的平均相似度作为一个特征,将该类型特征向量定义为:

$$V_s=<f_r> \tag{21}$$

3.4 特征集合

面对一个未使用HTTPS的网站,将上述特征组合构成总的特征集合 V。

面对一个使用HTTPS的网站,先使用ssl模块对数字证书进行解析。若规则匹配阶段未发现异常,则针对该URL进行URL特征和PageRank特征提取。并使用域名进行Whois和DNS信息进行提取,使用requests模块获得未加密的网页源代码文本。再进行网页标签,子链接和文本关键字和网页文本类钓鱼词汇特征提取工作。对网页内容中出现的URL,除对它们进行子链接特征提取工作外不进行其他特征提取工作。最终构成特征集合 V。

$$V=<V_u,V_p,V_W,V_D,V_H,V_{son},V_s,V_S> \tag{22}$$

3.5 分类方法

Nmap-RF一共选择了6种分类算法作为待选择算法,它们的优缺点见表2。其中朴素贝叶斯、支持向量机和决策树都属于数据挖掘十大经典算法。逻辑回归和多层感知机也是常见的分类方法,通过第4章的实验,最终确定选择了以决策树为基学习器的随机森林算法作为最终的分类方法。

4 实验

4.1 实验环境和数据来源

本文实验使用Windows系统,8 GB RAM的主机,代码全部使用Python开发,分类算法模型的构建使用开源的Scikit-learn。本文的数据集包括两个部分,一部分来自PhishTank的非法数据集,共计8 292个。一部分来自于从Yahoo收集得到的合法网页数据集,共计2 683个。

4.2 实验设计

本文使用了机器学习领域中常用的七折交叉验证法进行训练以及测试。实验部分首先检测三条针对使用了HTTPS的网站的过滤效果。第二步比较了多种常见分类算法,选择分类效果最佳和速度较快的分类器。第三步通过比较五种字符串相似度的算法,综合运行时间和准确率,挑选出合适的字符串相似度比较算法。第四步通过对子树数目参数的调整使RF分类器达到最优的分类性能。第五步通过更改TF-IDF的阈值检测对分类器性能的影响。第六步检测Whois特征对分类器的影响,最后将Nmap-RF方法从三个指标和其他方法做出比较。

本文选定同样使用集成方法的检测方法作为比较对象。庄蔚蔚[11]等人提出的SVMNB系统以改进支持向量机和拓展贝叶斯分类器为基础分类器。冯庆[12]等人提出的IPDWS系统通过对网页URL,网页链接以及页面文本信息进行特征提取。丁岩[15]通过规则匹配和逻辑回归结合建立分类器,与前两文不同的是,本文提出了三条规则对网页进行预过滤,满足了实时监测的要求。此外,本文以决策树为基分类器构造集成分类器。与丁岩[15]不同的是,本文规则针对HTTPS提出,且扩大了特征集合,使用了集成学习方法。在特征提取方面,本文不仅提取了URL字符特征,还对网页的DNS、Whois信息、网页文本、子链接以及PageRank等特征进行了提取。针对以上三种检测方法,本文设置了对比实验进行分析,表3和表4是训练集和测试集样本设置的数目。其中,Phish数据集共有130个使用了HTTPS的网站,Yahoo共有432个使用了HTTPS的网站。

表2 候选分类方法介绍

分类方法	优点	缺点
LR	计算代价低,易实现	容易产生欠拟合,分类精度不高
SVM	适用于小样本问题,泛化能力较强,较好地处理高维数据	解释性强,训练速度慢
NB	适用于大规模数据集,对结果解释容易理解	依赖于样本属性独立性假设
MLP	分类精度高,学习能力强;容错性高,具有更高的鲁棒性	参数较多,学习过程是黑盒的
DT	具有很好的可解释性,测试数据集时速度快	易出现过拟合,较难处理缺失数据
RF	不需要归一化处理就能处理数据,泛化性能更好	噪声较大时易过拟合

表3　训练集设置

数据源	PhishTank	Yahoo
样本数	2 000	6 000

表4　测试集设置

数据源	PhishTank	Yahoo
样本数	683	2 292

4.3　评价指标

在本章设置的实验中,主要采用准确率、召回率、F-Score作为检测过滤效果的主要评价指标。表5列出了被正确分类和错误分类的网页标识,并依此给出了准确率、召回率及F-Score的计算公式。

表5　分类混淆矩阵

	判定为钓鱼网页 P'	判定为合法网页 L'
钓鱼网页 P	$N_{P\to P'}$	$N_{P\to L'}$
合法网页 L	$N_{L\to P'}$	$N_{L\to L'}$

准确率的计算公式为:

$$T=\frac{N_{P\to P'}+N_{L\to L'}}{P+L} \tag{23}$$

召回率的计算公式为:

$$R=\frac{N_{P\to P'}}{P} \tag{24}$$

F-Score的计算公式为:

$$F=\frac{2TR}{T+R} \tag{25}$$

4.4　实验结果

4.4.1　规则匹配实验结果

首先对规则匹配阶段进行实验,由表6可知,在本文所使用的数据集中,一共有562个使用了HTTPS协议的网站。使用证书信息进行对应的规则匹配,成功过滤了14.6%的非法网页。

表6　基于规则匹配的检测效果

数据集合	PhishTank	Yahoo
使用HTTPS协议的网站	130	432
判定为钓鱼URL的数量	19	0

与IPDWS和SVMNB方法相比,增加了规则匹配阶段,减少了特征提取阶段,节约了加载模型和计算的时间,满足了实时检验的要求。与RMLR相比,规则更加简单,且对使用HTTPS的网站更有针对性。

4.4.2　不同分类方法选择

分类器检测精度取值为五次交叉验证的平均值,由图3和表7可知,结合F1-Score和准确率以及召回率来看,RF分类器更适用于钓鱼网站分类问题。

在对文本信息进行处理时,文本比较的算法往往对所用时间起重要影响。由图4可知,通过对比Jaccard距离、编辑距离、莱文斯坦比、Jaro距离以及Jaro_winkler距离这五种计算字符串相似度的方法,本文权衡了准确率和所用时间,最终选择了编辑距离作为文本相似的比较算法。

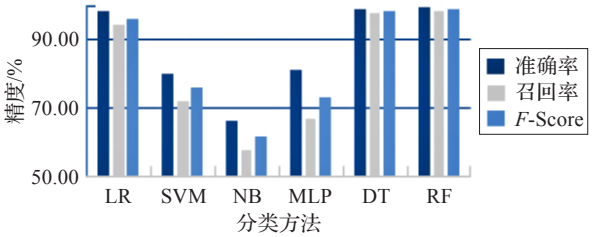
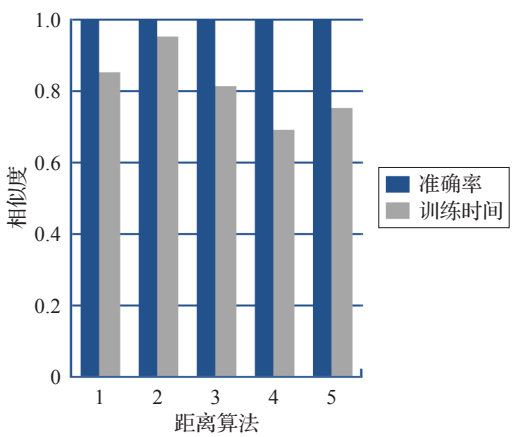


图3　不同分类器检测方法比较

表7　分类器检测结果比较

分类方法	LR	SVM	NB	MLP	DT	RF
准确率	0.983	0.801	0.662	0.810	0.988	0.995
召回率	0.946	0.722	0.579	0.672	0.979	0.985
F-Score	0.964	0.759	0.617	0.735	0.983	0.990



注:1—Jaro_winkler距离;2—Jacro距离;3—莱文斯坦比;4—编辑距离;5—Jaccord距离

图4　文本距离相似度方法比较

4.4.3　其他实验

对RF分类算法的参数进行调节,改变不同的子树从而得到最好的效果,由表8可知,综合效率和准确率两个因素,子树数目为50时最适合作为分类方法。

表8　RF分类器调参结果比较

序号	子树数目	准确率/%	时间/s
1	10	99.3	0.189
2	20	99.2	0.357
3	30	99.3	0.513
4	50	99.5	0.921
5	70	99.4	1.245
6	80	99.4	1.357
7	100	99.5	1.687

由表9可知,在权衡提取速度和准确值后使用0.5作为阈值对文本关键字提取可以保证提取速度和准确率达到良好的效果。

考虑到数据集中约有五分之一URL的Whois应答无内容,为此增设了Whois特征的实验,由图5可知,在增加了Whois特征后,召回率上升了约0.5%,准确率上

表9　TF-IDF调整阈值取值结果比较

阈值取值	词汇个数	提取速度/s	准确率/%
0.1	934	21.866	99.1
0.2	829	20.234	99.0
0.3	699	19.613	99.2
0.4	532	19.538	99.4
0.5	394	19.234	99.6
0.6	280	19.661	99.5
0.7	186	19.121	99.5
0.8	96	18.673	99.4
0.9	73	18.249	99.3

升了约0.2%。因此决定仍然保留Whois信息作为特征之一。表10是使用主成分分析法进行分析特征贡献度后得到的结果。可以看出,本文增加了网页文本和网页链接以及PageRank提供了约0.2的贡献率。

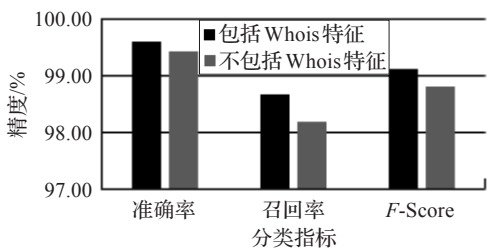


图5　Whois特征对结果的影响

表10　各个特征贡献率

特征类别	贡献度(约等于两位小数)
HTML特征	0.05
URL特征	0.41
Whois特征	0.11
DNS特征	0.22
Jaccord特征	0.01
PageRamk特征	0.11
文本信息特征	0.08
子链接特征	0.01

4.4.4　与其他方法的比较

综上所述,本文提出的随机森林分类器在子树为50的情况下,对指定特征记性提取训练,配合以过滤的预规则,所得到的分类器在分类正确率上可以达到99.6%。Nmap-RF从时间复杂度上看,特征提取中和DNS服务器,Whois服务器以及网页服务器交互所耗的时间主要较多,但由于都设置了timeout时间,因此总的消耗时间和条数呈线性关系。而模型对10 000条URL的识别时间总需要0.921 s,平均每条数据集的识别时间只需要9~10 μs。从空间复杂度上看,算法运行时内存占用小于1%,且中间的特征向量以矩阵的形式存放,只占很少的资源。与SVMNB和IPDWS以SVM和LR作为基学习器相比,以决策树作为基学习器的本方法在速度上更有优势。

由表11可知,在分类准确率、召回率和*F*-Score三个指标上,Nmap-RF优于上述三种方法。

表11　Nmap-RF与其他方法的比较　%

指标	SVMNB	IPDWS	RMLR	Nmap-RF
准确率	96.60	95.90	98.70	99.60
召回率	97.62	96.73	99.00	98.68
F-Score	97.11	96.30	98.80	99.13

5　结束语

面对日趋增长的钓鱼网站威胁,除了现有的对使用HTTP的网站进行研究,本文还针对使用了HTTPS的网站提出了一种新的Nmap-RF集成检测方法。首先根据现在数字证书的使用情况,提出三条简单有效的过滤规则。其次在网页基础信息(Whois、URL、DNS)上增加了对网页PageRank特征的提取,针对网页链接,使用类递归的形式进行处理,并增加了对网页文本关键字的处理。比对了多种计算字符串相似度的方法,选择了编辑距离作为文本比对方法。最后在多种常用分类方法中进行充分的比较,选择了在效率和准确率都表现优异的随机森林法。最终提出的检测方法准确率达到了99.6%。在未来,钓鱼攻击仍然将持续地威胁着互联网用户的财产安全,从研究的方面来说,还可以结合网页图像对网站合法性做出判断。

参考文献:

[1] 360互联网安全中心.2017中国网站安全形势分析报告[EB/OL].(2018-01-23)[2018-12-01].http://zt.360.cn/1101061855.php?dtid=1101062368&did=490995546.

[2] 360互联网安全中心.2018年上半年中国互联网安全[EB/OL].(2018-07-30)[2018-12-01].http://zt.360.cn/1101061855.php?dtid=1101062360&did=491357630.

[3] 中国反钓鱼网站联盟.2018年8月钓鱼网站处理简报[EB/OL].(2018-09-07)[2018-12-01].http://www.apac.cn/gzdt/201809/P020180918609574562007.pdf.

[4] Aleroud A,Zhou L.Phishing environments,techniques,and countermeasures:a survey[J].Computers & Security,2017,68:160-196.

[5] Hou Y T,Chang Y.Malicious web content detection by machine learning[J].Expert Systems with Applications,2010,37(1):55-60.

[6] Dunlop M,Groat S,Shelly D.GoldPhish:using images for content-based phishing analysis[C]//Fifth International Conferenceon Internet Monitoring,Spain,2010:123-128.

[7] Kang L C,Chang E H,Sze S N,et al.Utilisation of website logo for phishing detection[J].Computers & Security,2015,54:16-26.

[8] Li Y,Yang L,Ding J.A minimum enclosing ball-based support vector machine approach for detection of phishing websites[J].Optik-International Journal for Light and Electron Optics,2016,127(1):345-351.

(下转第200页)

[7] Shechtman E,Irani M.Matching local self-similarities across images and videos[C]//Proceedings of Computer Vision and Pattern Recognition,2007:1-8.
[8] Prisacariu V A,Timofte R,Zimmermann K,et al.Integrating object detection with 3D tracking towards a better driver assistance system[C]//Proceedings of International Conference on Pattern Recognition,2010:3344-3347.
[9] Douville P.Real-time classification of traffic signs[J].Real-time Imaging,2000,6(3):185-193.
[10] Cristianini N,Shawe-Taylor J.An introduction to support vector machines and other kernel-based learning methods[M].New York:Cambridge University Press,2001:1-28.
[11] Zaklouta F,Stanciulescu B,Hamdoun O.Traffic sign classification using K-d trees and random forests[C]//Proceedings of International Joint Conference on Neural Networks,2011:2151-2155.
[12] Romdhane N B,Mliki H,Hammami M.An improved traffic signs recognition and tracking method for driver assistance system[C]//Proceedings of IEEE International Conference on Computer and Information Science,2016:1-6.
[13] Zhang Y K,Hong C Y,Wang C.A real time rectangular speed limit sign recognition system[J].CAAI Transactions on Intelligent Systems,2010:6-16.
[14] 徐岩,韦镇余.一种改进的交通标志图像识别算法[J].激光与光电子学进展,2017,54(2):021001.
[15] Tang S,Huang L L.Traffic sign recognition using complementary features[C]//Proceedings of Asian Conference on Pattern Recognition,2014:210-214.
[16] Wu Y,Liu Y,Li J,et al.Traffic sign detection based on convolutional neural networks[C]//Proceedings of 2013 International Joint Conference on Neural Networks,2013:1-7.
[17] Sermanet P,Lecun Y.Traffic sign recognition with multi-scale convolutional networks[C]//Proceedings of International Joint Conference on Neural Networks,2012:2809-2813.
[18] 戈侠,于凤芹,陈莹.基于分块自适应融合特征的交通标志识别[J].计算机工程与应用,2017,53(3):188-192.
[19] 金戈,叶飞跃.基于LBP和HOG特征融合的行人检测[J].数字化用户,2017,23(40):186-187.
[20] Weijer J V D,Schmid C.Applying color names to image description[C]//Proceedings of IEEE International Conference on Image Processing,2007:493-496.
[21] Zhang J,Marszałek M,Lazebnik S,et al.Local features and kernels for classification of texture and object categories:a comprehensive study[C]//Proceedings of Computer Vision and Pattern Recognition Workshop,2006:213-238.
[22] Liu Y,Zheng Y F.One-against-all multi-class SVM classification using reliability measures[C]//Proceedings of IEEE International Joint Conference on Neural Networks,2005:849-854.
[23] Stallkamp J,Schlipsing M,Salmen J,et al.Man vs.computer:benchmarking machine learning algorithms for traffic sign recognition[J].Neural Networks,2012,32(2):323-332.
[24] Timofte R,Zimmermann K,Gool L V.Multi-view traffic sign detection,recognition,and 3D localisation[J].Machine Vision & Applications,2014,25(3):633-647.
[25] Zhang H,Bo W,Zheng Z,et al.A novel detection and recognition system for Chinese traffic signs[C]//Proceedings of IEEE Control Conference,2013:8102-8107.

(上接第88页)

[9] Hu Z,Chiong R,Pranata I,et al.Identifying malicious web domains using machine learning techniques with online credibility and performance data[C]//2016 IEEE Congress on Evolutionary Computation,2016:5186-5194.
[10] Martínez S J,Pla F.Supervised feature selection by clustering using conditional mutual information-based distances[J].Pattern Recognition,2010,43(6):2068-2081.
[11] 庄蔚蔚,叶艳芳,李涛,等.基于分类集成的钓鱼网站智能检测系统[J].系统工程理论与实践,2011(10):2009-2010.
[12] 冯庆,连一峰,张颖君.基于集成学习的钓鱼网页深度检测系统[J].计算机系统应用,2016(25):47-55.
[13] Dong Zheng,Kane K,Camp L J.Detection of rogue certificates from trusted certificate authorities using deep neural networks[J].ACM Transactions on Privacy and Security,2016,19(2):5.
[14] Clark J,Van Oorschot P C.Revisiting past challenges and evaluating certificate trust model enhancements[C]//IEEE Symposium on Security and Privacy,2013:511-525.
[15] 丁岩,努尔布力.基于URL混淆技术识别的钓鱼网页检测方法[J].计算机工程与应用,2017,53(20):75-81.
[16] 丁岩.基于机器学习的钓鱼网页检测方法研究[D].乌鲁木齐:新疆大学,2018.
[17] Ma J,Saul L K,Savage S,et al.Learning to detect malicious URLs[J].ACM Transactions on Intelligent Systems & Technology,2010,2(3):493-500.
[18] Le A,Markopoulou A,Fabloutsos M,et al.PhishDef:URL names say it all[J].Computer Science,2010,28(6):191-195.
[19] Zhang Y,Hong J I,Cranor L F.Cranor.Cantina:a content-based approach to detectingphishing web sites[C]//Proceedings of the 16th International Conference on World Wide Web,2007:639-648.