基于机器学习方法的钓鱼网址识别的研究

叶飞[1]，龚俭[1]，杨望[1]

（1. 东南大学计算机科学与工程学院，南京，211189）

摘　要：钓鱼式网络攻击是在线经济犯罪的主要形式之一。本文利用 URL 的文本特征对于识别和分类钓鱼网站进行研究，使用朴素贝叶斯，逻辑回归，径向量核函数支持向量机以及线性支持向量机作为分类器进行实验，实现了网络钓鱼检测,并且可以保证检测的效率和精度。实验结果证明，采用 URL 文本特征和机器学习方法的网络钓鱼检测系统能达到较高的检测精度。

关键词：钓鱼网站;入侵检测;机器学习;文本分类;支持向量机

Study on Phishing URL Detection Based on Machine Learning Method

Fei Ye[1], Jian Gong[1], Wang Yang[1]

(1. School of Computer Science and Engineering, Southeast University, Nanjing, 211189)

Abstract: Online phishing attack is one of the main approach of economic crime. In this paper, we proposed a method using text features of URLs for the identification and classification of phishing website. This method utilize naive bayes, logistic regression, radial basis function support vector machine (RBF-SVM) and linear support vector machine (Linear-SVM) as classifier to implements the phishing detection system and guarantee the efficiency and accuracy of detection. The results show that machine learning method for detection phishing websites with text features of URLs can achieve higher precision.

Key words: Phishing Websites; Intrusion Detection; Machine Learning; Text Classfication; Support Vector Machine

近年来，互联网在中国得到了蓬勃的发展，但越来越严重的安全问题引起了人们的担扰，其中，网络钓鱼（Phishing）就是网络欺诈的代表。网络钓鱼式攻击是一种在线身份伪造的欺诈方式，使用社会工程学和技术伪装等攻击手段骗取用户的信任，窃取用户的个人身份数据和金融账号等敏感信息。金山网络发布的《2012 年度计算机病毒及钓鱼网站统计报告》显示，2012 年全球新增钓鱼网站 163.1 万个，而这个数字在 2011 年为 45 万个，钓鱼网站总量在一年里快速增长了 3.6 倍。根据中国反钓鱼网站联盟的年度报告，2012 年 1 月 1 日至 11 月 20 日，中国反钓鱼网站联盟共处理钓鱼网站 24535 个，涉及淘宝网、工商银行、央视、腾讯公司四家单位的钓鱼网站总量占全部举报量的 80.09%。其中，淘宝网处于钓鱼网站仿冒对象的第一位。这充分反映了网络钓鱼对网络和个人信息和经济安全的严重威胁。静态地来看，钓鱼式网络攻击实际上是一些伪装成正规网站的 phishing Web 网页。通过对这些 Web 网页及其 URL 研究表明，它们存在一些明显的特征。如果能把这些 phishing 网址 URL 或 Web 网页识别出来，能对钓鱼式网络攻击进行防范和打击。

因此，如果可以预先告知使用者当前的网址是有危害的，便可以减轻此攻击所造成的伤害。最常使用的机制是黑名单(blacklist)，可以加附在浏览器或搜寻引擎上，而黑名单的资料收集是由手动回报、honeypot 和网络筛选后的结果而来。例如 PhishTank[1]通过用户举报和人工审核的方式，建立了可信度最高的钓鱼式网络攻击的 URL blacklist 供浏览器插件过滤掉这些已报告的攻击网站。除了黑名单,目前有的解决方案是去分析网页内容，例如根

作者简介：叶飞，（1989-），男，硕士研究生，E-mail：fye@njnet.edu.cn；龚俭，（1957-），男，教授、博士生导师，E-mail：jgong@njnet.edu.cn；杨望，（1979-），男，讲师，E-mail：wyang@njnet.edu.cn

据 Web 页面的异常特征进行分析[2]，或者使用基于嵌套 E M D (Nested Earth M over's Distance)的网页相似度判定算法进行钓鱼网页检测[3]，但这些方法会增加处理时间。也有采用仅对 URL 在线识别的检测方法[4]，通过已知攻击特征等先验知识进行学习，对 URL 进行分类检测。但是有很多恶意网址是过于新的，无法即时列在黑名单内。

本文使用轻量的 URL 分类方法，仅仅使用 URL 的文本特征来区分用网址是否用于钓鱼，并以黑名单为基础，用来预测未列在黑名单上的新网址，较大地提升了判断的效率。本文第 1 节通过对钓鱼 URL 实例的分析，提出了基于文本特征分析的钓鱼检测技术。第 2 节对检测关键技术给出了具体说明。第 3 节介绍了所用的数据集和来源。第 4 节通过实验对本方法的检测能力进行验证。第 5 点为最后一节，对全文进行总结并提出下一步的研究重点。

1 钓鱼 URL 实例分析

攻击者为了让用户访问其建立的钓鱼网站，往往会使用特殊的 URL 地址来欺骗用户。在本节中，我们通过分析三个钓鱼 URL 实例来为下文的特征提取做准备。如图 1 所示为某个模仿中国银行网上银行 http://www.boc.cn/ 的钓鱼 URL 地址。该钓鱼 URL 使用二级域名来伪装自己，其 hostname（主机名）中有较多的 tokens（被符号隔开的字符串）和“.”，并且长度较长。图二中的两个钓鱼 URL 分别使用了免费主域名.tk 和不常见的顶级域名.co.za。这些文本上的特征都可以作为对于 URL 进行分类的依据。

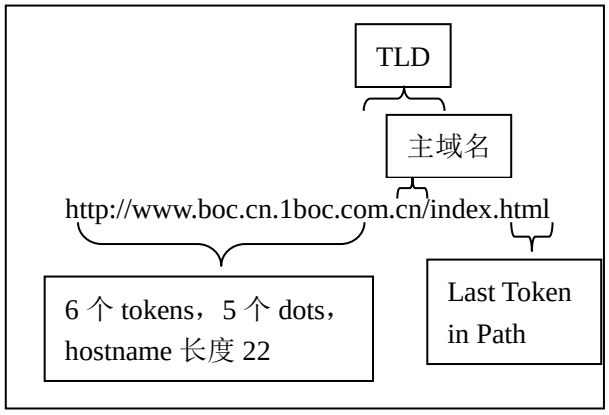

图 1 国内网银钓鱼 URL 示例

本文提出一种基于 URL 文本特征分类的钓鱼网站检测技术。根据钓鱼 URL 为了迷惑用户而形成的视觉特征进行分析总结，找到钓鱼 URL 的一般规律性特点，作为分类测度来训练机器学习分类器，从而建立分类模型，对于新的未知钓鱼 URL 进行世识别。

> http://us.bsgenet.tk/login.asp?ref=https%3A%2F%2Fus.battle.net%2Faccount%2Fmanagement%2Findex.xml&app=bam
>
> http://barclaysbank.co.uk.olb.auth.loginlinkaction.jscript32dsg234235gsf23423423gfgd2342342.jsessions324234j23hgrjhfjfgj.junglerumblepartyvenue.co.za/bank.barclays.co.uk.olb.auth.LoginLink.action/

图 2 国外钓鱼 URL 示例

2 检测关键技术

本文主要选取了 URL 中的特定文本特征，采用了朴素贝叶斯（Na ïve Bayes），RBF-SVM，线性 SVM 以及逻辑回归（Logistic Regression）作为学习算法。使用 WEKA 作为实验工具。

2.1 总体思路

本文使用二元的分类法将恶意的网址标记为正(positive)和良好的网址标记为负(negative)。用学习的(learning-based)方法，用多种特征去训练后，再去测试是否可以正确区分恶意的和良好的网址。本文依据 URL 词汇的特征来区分网址，而且完全不考虑网页的内容。尽管网页内容可以提升准确率，但本文不采用原因由以下四点：第一点，避免下载网页的内容。第二点，下载与分析网页内容的时间花费较高。第三点，使用网址特征可以让分类器应用在更多的含有网址的内容上。最后一点，能可靠的取得用来训练和测试的资料。

2.2 URL 特征选取

如果单纯地以 Web 页面为检测对象，对页面内容进行识别和保护，则大大降低了钓鱼网站检测的效率，毕竟钓鱼网站跟海量的正规网站相比还是占少数的，且其存活期很短。由于钓鱼骗术花样百出，因此就像杀毒软件不能查杀所有病毒一样，单纯地

针对页面特征的钓鱼网站检测也不一定能检查出具有新特征的钓鱼网站，同时还容易误伤正规网站，即存在较高的漏检率和误检率；而钓鱼网站的 URL 往往存在某些误导用户的特征，本文主要选取词汇特征作为测度。使用词汇特征是因为恶意网址通常会去欺骗使用者的眼睛。词汇特征便可以做为分类的依据。本文在总结前人研究的基础上[5]，挑选出以下最有可能是钓鱼式攻击网站的词汇特征，分为以下两个部分：

2.2.1 实数特征：

1) Length of Hostname。钓鱼网站往往使用临时申请的随机长域名;有些钓鱼网站为了诱人上钩，经常包含被模仿网站的域名,因此导致主机名较长。

2) Length of URL。很多钓鱼 URL 都需要经过跳转，因此包含掉转后的地址；或者有意使用很长的 URL 是用户失去阅读的耐心。

3) Hostname 中“.”的个数。检测钓鱼 URL 是否使用了欺骗性的二级域名。一般用户只会阅读 URL 的前几个字段，由此攻击者会使用二级域名来进行钓鱼。经过实现表明主机名中“.”的个数越多，越有可能为钓鱼 URL。

4) URL 中各类分割 token 的数目(.-_?=&)。钓鱼网站较多使用这些分隔符来迷惑用户。

2.2.2 二进制特征：

由于有些域名或者单词常常被攻击者滥用，本文采用词袋模型（bag of words），对于下面提到所有 token 生成一串二进制（0，1）向量，为后面的分类做准备。本文需要的词袋来自以下五种 token：

1) TLD。
2) Tokens in Primary Domain。
3) Tokens in Host Name。
4) Tokens in Path
5) Last Tokens of Path

2.3 分类方法

URL 分类的本质是文本分类。目前较为著名的文本分类方法[6][7]有 Bayes、SVM、KNN、决策树、逻辑回归等。本文在总结前人研究的基础上，选择使用朴素贝叶斯、SVM（RFB 和线性）和逻辑回归进行分类实验。

贝叶斯（Bayes）分类方法是一种最常用的有指导的方法，以贝叶斯定理为理论基础，是一种在已知先验概率与条件概率的情况下的模式识别方法。朴素贝叶斯分类器（Naive Bayes)假设特征对于给定类的影响独立于其它特征，即特征独立性假设。 对 URL 分类来说 ,它假设各个 token 之间两两独立。本文选用朴素贝叶斯方法进行试验。

Logistic 回归是估计后验概率的一种简单而有效的监督学习算法，它不需要任何关于样本的先验知识和分布假设，对自变量类型也不做要求，其结果易于解释，近几年来被广泛应用于机器学习、数据挖掘和模式识别等领域。与朴素贝叶斯模型是在事先知道先验概率且各类指标服从多元正态分布的前提下用判别函数法估计参数不同，Logistic 回归是直接根据多元指标通过回归模型求出某一事件发生的估计概率，并根据概率大小判别归类。

SVM 是 Vapnik 等于 1995 年提出的，它在解决小样本、非线性及高维模式识别中表现出许多特有的优势。SVM 方法是通过一个非线性映射 p，把样本空间映射到一个高维乃至无穷维的特征空间中（Hilbert 空间），使得在原来的样本空间中非线性可分的问题转化为在特征空间中的线性可分的问题。一般的升维都会带来计算的复杂化，SVM 方法巧妙地解决了这个难题：应用核函数的展开定理，就不需要知道非线性映射的显式表达式；由于是在高维特征空间中建立线性学习机，所以与线性模型相比，不但几乎不增加计算的复杂性，而且在某种程度上避免了“维数灾难”。选择不同的核函数，可以生成不同的 SVM，本文选用了径向基函数（RBF）以及线性核函数进行试验。

3 数据集

本文使用 URL 敏感特征进行检测分类,实验所用的正例样本白名单是通过访问 http://random.yahoo.com/bin/ry，从而得到合法正规网站。而反例样本黑名单则是通过 PhishTank 提供的黑名单数据库获取的。本次实验采用了 2012 年 5 月 24 日 PhishTank.com 提供的 phishing URL 黑名单更新数据。本文选取了 3653 个站点信息用作正例样本，而反例样本使用了 2700 个钓鱼 URL。在敏感特征的选取方面，采用简单而功能强大的 Python 脚本语言实现，本次实验中使用词袋（bag of words）模型，得到的样本为 10716 维的向量。

4 实验结果及分析

本文使用常用的数据挖掘软件 WEKA[8]进行试验，所有分类方法采用默认参数，随机抽取 2/3 数据作为训练数据，另外 1/3 数据用来测试，检测钓鱼 URL 的实验结果如表 1、图 3、表 2 所示：

表 1 机器学习算法检测试验效果

	准确率	误报率	漏报率
贝叶斯	76.50%	6.20%	33.70%
RBF-SVM	93.20%	10.00%	4.40%
逻辑回归	62%	51%	0.60%
线性 SVM	93.90%	6.90%	4.60%

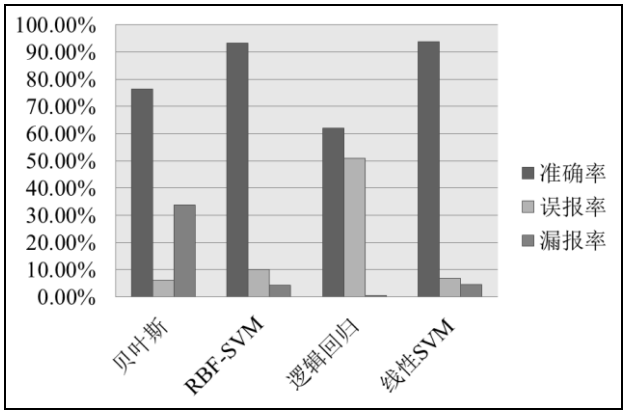

图 3 机器学习检测结果

根据实验结果显示，相对于其他三个分类器，线性 SVM 的平均性能最佳。准确率方面线性 SVM 和 RBF-SVM 都超过了 90%的分类器，而其误报率和漏报率也在可以接受的水平。上述结果同样可以看到，在分类准确率方面，各个分类器的差距并不明显。但是在误报率上，逻辑回归超过了 50%，远远超过了其他分类器。类似的结果也发生在漏报率上，朴素贝叶斯的漏报率也明显过高。

表 2 机器学习算法训练时间

	贝叶斯	RBF-SVM	逻辑回归	线性 SVM
训练时间	38.12s	17.14s	161.66s	3.44s

从训练时间来看，朴素贝叶斯，RBF-SVM 以及线性 SVM 都在可以接受的范围，而逻辑回归的训练时间很长，如果加上交叉验证，训练时间还将进一步增长，但是分类效果却不是很好。值得一提的是，线性 SVM 的训练时间最短，只有 3.44 秒，仅为逻辑回归的 2%。

5 总结

本文提出了以网址为基础了分析方法，使用网址的各种特征及机器学习的方式来区分恶意或良性的网址。文中比较了不同分类器的差异，得出了线性 SVM 分类效果较好的结论。在后续的实验中，使用开源的 liblinear[9][10]库编写代码，对于 35,000 个白名单和 5,000 个黑名单做实验的结果显示，分类准确率达到了 99.2419%，而误报率不到 1%。可见取得了良好的效果。下一步将在算法的实现上重点改进 SVM 算法中核函数的选择、参数的优化及其在实际网络环境中的适应性。

参考文献

[1] OpenDNS.PhishTank. http://www.phishtank.com.

[2] Basnet R, Mukkamala S, Sung A H. Detection of phishing attacks: A machine learning approach[M]//Soft Computing Applications in Industry. Springer Berlin Heidelberg, 2008: 373-383.

[3] 曹玖新,毛波,罗军舟,刘波.基于嵌套 EMD 的钓鱼网页检测算法[J].计算机学报,2009,32(5):922-929.

[4] Justin M K. Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs[C]//Proc. of ACM SIGKDD'09. Paris, France: ACM Press, 2009: 1245-1253.

[5] 何高辉,邹福泰,谭大礼,王明政. 基于 SVM 主动学习算法的网络钓鱼检测系统[J]. 计算机工程,2011.10.

[6] Tom M.Mitchell. 机器学习[M]. 北京:机械工业出版社,2009.

[7] Richard O,Duda. 模式分类[M]. 北京:机械工业出版社,2007.

[8] Hall M, Frank E, Holmes G, et al. The WEKA data mining software: an update[J]. ACM SIGKDD Explorations Newsletter, 2009, 11(1): 10-18.

[9] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A Library for Large Linear Classification, Journal of Machine Learning Research 9(2008), 1871-1874.

[10] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011.