doi :10.3969/j.issn.1671-1122.2012.01.006

基于异常特征的钓鱼网站 URL 检测技术

黄华军，钱亮，王耀钧

（中南林业科技大学计算机与信息工程学院，湖南长沙 410004）

摘 要：典型的网络钓鱼是采用群发垃圾邮件，欺骗用户点击钓鱼网站 URL 地址，登录并输入个人机密信息的一种攻击手段。文章通过分析钓鱼网站 URL 地址的结构和词汇特征，提出一种基于异常特征的钓鱼网站 URL 检测方法。抽取钓鱼网站 URL 地址中 4 个结构特征、8 个词汇特征，组成 12 个特征的特征向量，用 SVM 进行训练和分类。对 PhishTank 上 7291 条钓鱼网站 URL 分类实验，检测出 7134 条钓鱼网站 URL，准确率达到 97.85%。

关键词：网络钓鱼；钓鱼网站 URL；支持向量机；特征向量

中图分类号：TP393.08 **文献标识码：**A **文章编号：**1671-1122（2012）01-0023-03

Detection of Phishing URL Based on Abnormal Feature

HUANG Hua-Jun, QIAN Liang, WANG Yao-Jun

(*Department of computer and information engineering Central South University of Forestry and Technology, Changsha Hunan* 410004*, China*)

Abstract: Phishing tries to lure her victim into clicking a phishing URL pointing to a spoof page via spam-email to harvest financial information. In this paper, a novel method is proposed to detect phishing URL based on abnormal feature. The feature vector is constructed with 12 features to model the SVM, which 4 features are the structure feature of the phishing URL, 8 features are lexical feature. The method can correct to classify 7134 phishing URLs of 7291downloaded in PhishTank achieve, and the correct ratio of detection is 97.85%.

Key words: Phishing; Phishing URL; SVM; feature vector

0 引言

网络钓鱼（Phishing）是基于社会工程学的一种攻击手段[1, 2]。它通过垃圾邮件、即时聊天工具、手机短信或网页虚假广告发送声称来自于银行或其他知名机构的欺骗性信息，意图引诱用户登录看起来极其真实的假冒网站，给出敏感信息（如用户名、口令、账号 ID、ATM PIN 码、信用卡）的一种攻击方式。最典型的网络钓鱼攻击将用户引诱到一个通过精心设计与目标组织的网站非常相似的钓鱼网站上，并获取用户在该网站上输入的个人敏感信息。

网络钓鱼防御是网络钓鱼的对抗技术，现主要集中在钓鱼网站检测[3]、垃圾邮件过滤[4]、钓鱼网站追踪[5]、网络钓鱼行为分析[6]、终止钓鱼网站域名解析[7]等多个研究领域，以钓鱼网站检测最活跃。钓鱼网站检测是在浏览器中安装检测插件，当用户浏览可疑钓鱼网站时，浏览器的插件将提醒用户当前网站为钓鱼网站。插件的检测算法采用 URL 检测技术[8-17]、基于启发式检测技术[18]、基于视觉相似检测技术[19]等判别钓鱼网站。

URL 检测技术直接分析钓鱼网站 URL 地址，判断是否为钓鱼网站链接，具备不需要下载钓鱼网站的网页内容进行分析，检测性能高等优点。本文在已有相关文献基础上，分析的钓鱼网站 URL 地址的结构特征和词汇特征，选取 4 个结构特征、8 个词汇特征，构建特征向量，使用 libSVM[20] 作为分类的模型。通过对 PhishTank 上 7291 条钓鱼网站 URL 分类实验，检测出 7134 条钓鱼网站 URL，准确率达到 97.85%。

收稿时间：2011-12-15

基金项目：国家自然科学基金项目［61073191］、湖南省自然科学基金资助项目 [10JJ4043,10JJ5062]、湖南省教育厅资助项目 [08B091]、湖南省科技重大专项项目 [2010J05]、湖南省科技计划重点项目 [2010NK2003]、湖南省科技计划项目 [2010TZ4012]

作者简介：黄华军（1978-），男，湖南，硕士生导师，副教授，中国计算机学会会员，博士，主要研究方向：网络与信息安全、网页信息隐藏、网络钓鱼检测；钱亮（1985-），男，安徽，硕士研究生，主要研究方向：网络钓鱼检测；王耀钧（1976-），男，湖南，硕士研究生，主要研究方向：网络钓鱼检测。

1 相关研究工作

1.1 基于URL黑名单检测技术[8-11]

基于URL检测是指利用URL地址，判断当前的网站是否为钓鱼网站。最初的方法是利用黑名单中存储被确认的钓鱼网站URL地址，当浏览器浏览时，提醒用户当前网站为钓鱼网站。Microsoft IE、Google Safe Browser、Netcraft Tool Bar、eBay Tool Bar、McAfee SiteAdvisor等知名IT企业采用黑名单防御钓鱼网站。为验证黑名单性能，先后收集了三周内10000条钓鱼网站的URL地址，测试微软IE浏览器和谷歌安全浏览器的检测性能。经分析发现，Google能够识别90%的钓鱼网站URL地址[8]。Sheng等人使用钓鱼网站URL，测试8种网络钓鱼防御工具中黑名单更新速度。他们发现小于20%的网络钓鱼防御工具能在短时间内（Hour Zero）识别钓鱼网站。尽管URL黑名单检测技术简单、检测率精确，但存在无法检测不在黑名单内的钓鱼网站，且确认黑名单需要人工验证费时耗力的问题[9]。

1.2 基于机器学习的检测技术[12-17]

基于机器学习的URL检测技术是直接利用URL检测钓鱼网站，主要流程如下：选择钓鱼网站URL特征向量，生成训练数据，训练构建分类器模型，应用分类器分类URL。此类检测特征选取和分类器构建是关键。

1）Garera算法[12]。Garera等人分析钓鱼网站URL结构，详细介绍特征集合选取过程，利用回归滤波器（Logistic Regression Filter）分类URL。对钓鱼网站URL结构进行分析，得出4种类型的URL结构。特征集合由页面特征、域名特征、类型特征和单词特征共18个特征构成。页面特征借助谷歌搜索引擎，选取URL页面排名（Page Rank of URL）、域名页面排名（Page Rank of Host）、页面排名存在爬行数据库（Page Rank Present in Crawl Database）、页面存在索引数据库（Page Present in Index）、两个页面质量评价（Page Quality Score）共6个特征；域名特征选择域名在白名单表1个特征；类型特征选择类型I、II、和III，3个特征；单词特征选择secure、account、webscr、login、ebayiapi、signin、banking和confirm 8个单词。

2）Ma算法[14-16]。与Garera等人采用18个特征作为分类网络钓鱼URL不同的是，Ma等人分析可疑URL的词汇（Lexical Features）和主机属性（Host-Based Features），采用词袋模型（Bag-of-Words）表示特征，获得了成千上万的特征。在词汇特征中，一方面考虑主机名长度、URL长度、URL中点号数等；另一方面，对于URL中主机和路径中每一个词汇符号，采用词袋模型建立一个二值特征。在主机特征中，考虑了IP地址属性、WHOIS属性、域名属性和地理位置属性。

考虑到批量学习（Batch Learning）和在线学习（Online Learning）性能要求，对于批量学习，Ma等人分析了朴素贝叶斯（Naïve Bayes）、支持向量机（SVM）和回归滤波（Logistic Regression）的分类性能[17]；在线学习中，研究了感知器（Perceptron）、随机梯度下降回归滤波（Logistic Regression with Stochastic Gradient Descent）、被动贪婪算法（Passive-Aggressive Algorithm）和秘密权证算法（Confidence-Weighted Algorithm）的分类性能。

与Ma相类似的算法还包括Blum等人提出的基于词汇特征的钓鱼URL的分类算法，详细分析参见文献[17]。

2 检测流程与特征向量

2.1 检测流程

McGrath等人经过详细实验，分析了钓鱼者做法，剖析钓鱼URL结构和域名、钓鱼网站域名注册信息、域名注册到使用的时间、钓鱼网站主机以及钓鱼网站生存周期[6]。他们的实验分析对检测钓鱼URL起到前提依据。分析Garera算法和Ma算法的特点和PhishTank中确认的钓鱼URL地址，基于SVM，设计一种钓鱼网站URL检测技术。

图1是钓鱼网站URL检测的基本流程，包括以下步骤：1）获取大量的测试URL并且确定URL是否为钓鱼网站URL地址；2）抽取URL的12个特征向量；3）在抽取完特征值后，生成训练样本，训练SVM；4）对可疑URL，生成判别的特征向量，用SVM进行分类检测，以最终判断是否为钓鱼网站的URL地址。

图1 钓鱼网站URL检测流程图

2.2 特征向量

特征向量关系到分类的性能。通过对已有钓鱼网站URL和相关文献分析的基础上，选取钓鱼网站URL地址的4个结构特征、8个词汇特征，共12个特征组成特征向量FV:

$$FV=<F_1,F_2,F_3,F_4,F_5,F_6,F_7,F_8,F_9,F_{10},F_{11},F_{12}>\cdots\cdots\cdots\cdots(1)$$

F_1：URL为IP地址；F_2：URL中存在@符号；F_3：URL长度超过23个字符；F_4：URL域名长度不超过7个字符，F_5：URL中存在secure；F_6：URL中存在webscr；F_7：URL中存在account；F_8：URL中存在login；F_9：URL中存在ebayisapi；F_{10}：URL中存在signin；F_{11}：URL中存在banking；F_{12}：URL中存在confirm。

特征F_1到F_4是钓鱼网站URL的结构特征，图2给出了

几种常见的钓鱼网站URL结构特点。特征 F_5 到 F_{12} 是钓鱼网站URL的词汇特征，文献[12]给出了8个特征在白名单和黑名单中出现的比例，见表1。

表1 钓鱼网站URL词汇特征(%)

features	ratio in whitelist	ratio in blacklist	features	ratio in whitelist	ratio in blacklist
confirm	0.23	4.25	ebayisapi	1.5	13.9
account	1.5	4.9	webscr	0.32	14.2
banking	0.87	7.95	login	2.61	21.53
secure	0.16	9.88	signin	0.95	23.29

在抽取这12个特征值以后并赋值，其中赋“1”代表为钓鱼网站；赋“0”代表为非钓鱼网站。结构特征 F_1 到 F_4 如下所示：

$$F_1=\begin{cases}1, & \text{if URL is IP address}\\ 0, & \text{others}\end{cases} \quad (2)$$

$$F_2=\begin{cases}1, & \text{if URL contain @}\\ 0, & \text{others}\end{cases} \quad (3)$$

$$F_3=\begin{cases}1, & \text{if length(URL)>23}\\ 0, & \text{others}\end{cases} \quad (4)$$

$$F_4=\begin{cases}1, & \text{if length(DN)<7}\\ 0, & \text{others}\end{cases} \quad (5)$$

公式4、公式5中Length函数是求字符长度，文献[6]给出钓鱼网站URL的长度特征。

对钓鱼网站URL的字符特征，采用公式6统一表示：

$$F_i=\begin{cases}1, & \text{URL contain w}\\ 0, & \text{others}\end{cases} \quad (6)$$

$$i\in\{5,6,7,8,9,10,11,12\},$$

$$w\in\{confirm, account, banking, secure, ebayisapi, webscr, login, signin\}$$

Phishing URL #1
http://210.80.154.30/~test3/.signin.ebay.com/ebayisapidllsignin.html
http://0xd3.0xe9.0x27.0x91:8080/.www.paypal.com/uk/login.html
Phishing URL #2
http://21photo.cn/https://cgi3.ca.ebay.com/eBayISAPI.dllSignIn.php
http://2-mad.com/hsbc.co.uk/index.html
Phishing URL #3
http://www.volksbank.de.custsupportref1007.dllconf.info/r1/vm/
http://sparkasse.de.redirector.webservices.aktuell.lasord.info
Phishing URL #4
http://www.wamuweb.com/IdentityManagement/
http://mujweb.cz/Cestovani/iom3/SignIn.html?r=7785

图2 钓鱼网站URL结构特征

3 实验结果

特征提取算法利用Java编程实现，分类算法利用libSVM。图3为一条URL特征值的抽取，当输入为http://www.baidu.com时，因为百度网址为非钓鱼网站，所以抽取的特征向量全为0。当输入的地址为http://210.80.154.30/的时候，特征向量中第一个值为1，如图4所示。图5是libSVM训练集的部分截图。

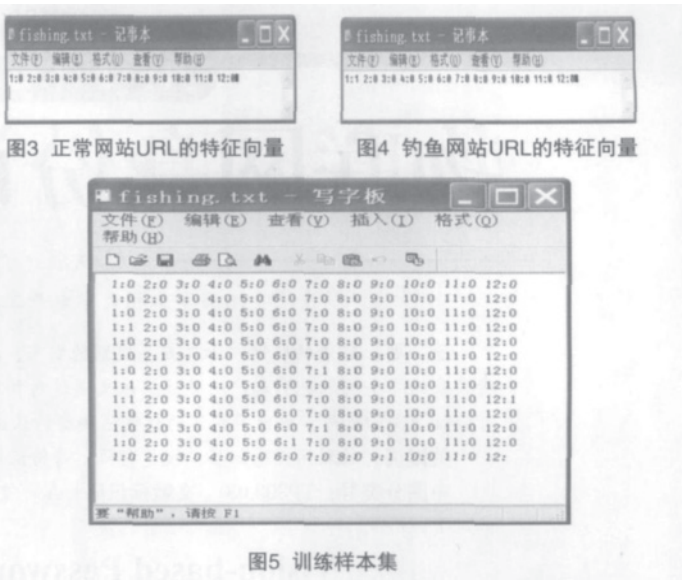

图3 正常网站URL的特征向量　图4 钓鱼网站URL的特征向量

图5 训练样本集

通过对PhishTank上7291条钓鱼网站URL分类实验，检测出7134条钓鱼网站URL，准确率达到97.85%。

4 结束语

网络钓鱼除了带来经济损失，同时也使网民对电子商务产生不信任心理，减少甚至避免使用某些网络应用，从而阻碍我国互联网的深入发展。网络钓鱼防御不仅仅是技术问题，它还涉及提高网民的安全防范意识和技能，商家、各种支付系统和Internet服务提供商要对用户负责任，有很好的配套措施，法律、政府安全部门尽快制订相关管理细则，加大打击与惩罚力度，共同营造一个安全、可信的互联网环境。

本文分析钓鱼网站URL地址的结构和词汇特征，构建12个特征的特征向量，采用libSVM进行分类，实验结果表明该方法能很好的检测出钓鱼网站URL地址，可应用在IE插件过滤钓鱼网站的URL地址。（责编 程斌）

参考文献：

[1] Anti-Phishing Working Group [EB/OL]. http://www.antiphishing.org, 2008-01/2011-12-15.

[2] PhishTank [EB/OL]. http://www.phishtank.com, 2011-04/2011-12-15.

[3] Engin Kirda, Christopher Kruegel. Protecting Users against Phishing Attacks[J]. The Computer Journal, 2006, 49(05):554-561.

[4] Ian Fette, Norman Sadeh, Anthony Tomasic. Learning to Detect Phishing Emails[C]. In Proc. of the WWW 2007, Alberta, Canada, May 8-12, 2007: 649-656.

[5] Chenfeng Vincent Zhou, Christopher Leckie, Shanika Karunasekera. Collaborative Detection of Fast Flux Phishing Domains[J]. Journal of Networks, 2009, 4(01):75-84.

[6] D. Kevin McGrath, Minaxi Gupta. Behind Phishing: An Examination of Phisher Modi Operandi[C]. In Proc. of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats, California USA, April 15 2008:1-8.

下转第67页

的关系，Mail capture 采用限制蜜罐系统同时连接数，周期性的允许定量的数据传出等方法来降低对系统的影响。

4.2 Honeyd部署位置的选取

Mail capture 配置的位置选取在防火墙的 DMZ 区，相对于防火墙外来说，这样可以有效与防火墙内部的系统进行数据交互并且处于相对安全的环境中，而相对于防火墙内部来说，这样会接受到更多的信息，因为防火墙的保护，内部的主机只能接受到很少的攻击信息，不利于蜜罐的正常工作。综上所述，最为合适的位置应该部署在防火墙的 DMZ 内。但部署在 DMZ 内的难度很大，一旦蜜罐被攻陷，攻击者会使用蜜罐作为跳板来攻击其他的服务，所以要使其他服务与蜜罐安全的隔离，这样也会增加 DMZ 部署的负担。

4.3 Honeyd采集数据的方式

Mail capture 所使用的数据采集方式为 ARP 欺骗方式，相对而言以 ARP 欺骗方式进行数据采集的蜜罐系统更容易配置在一个现有的网络之中。Honeyd 采集数据的方式大体上分为两种，第一种方法称之为 Blackholing，一般用于一个没有任何活动系统的完整网络，这意味着攻击者所瞄准的是特定网络中的任意 IP 地址，都可以视为攻击。因此这种方法就是将整个网络的流量都直接路由到 Honeyd 主机上；第二种方法为 ARP 欺骗，这种方法应用在一个网络中同时具有实际存在的系统和虚拟的系统，其目的在于将所有非存在的系统的流量都转发给 Honeyd 主机[8]。

5 结束语

蜜罐在安全领域的应用已经越来越广泛，相对于其他机制，蜜罐系统部署简单，配置灵活，收集到的数据有很大的针对价值，本文就是根据蜜罐技术模拟出 SMTP 的开放中继服务和开放代理服务，然后构建出一个用于收集垃圾邮件的系统，用于反网络钓鱼探测系统的研究。（责编 张岩）

参考文献：

[1] 中国反钓鱼网站联盟 . 2011 年 4 月钓鱼网站处理简报 [R]. 2011.4.

[2] 孙言 . 反网络钓鱼技术研究 [D]. 北京：中国人民公安大学，2009.

[3] 刘晓娟 . 基于邮件信息提取与分析的反网络钓鱼技术研究 [D]. 北京：中国人民公安大学，2010.

[4] 孙言、杜彦辉 反网络钓鱼中 UNICODE 字符相似度评估算法研究 [J]. 计算机工程与应用，2008，(26)：86-87.

[5] 翟继强，叶飞 . 利用 Honeyd 构建虚拟网络 [J]. 计算机安全，2006，(03)：46-48.

[6] Steding-Jessen, K., Vijaykumar, N. L., and Montes, A. Using low-interaction honeypots to study the abuse of open proxies to send spam[J] . INFOCOMP Journal of Computer Science 2008.

[7] Roshen Chandran, Sangita Pakala. Simulating Networks with Honeyd[EB /OL] . http://www. paladion. net/ papers/ simulating_networks_with _Honeyd.pdf, 2011-12-15.

[8] Lance Spitzner honeypots：追踪黑客 [M]. 北京：清华大学出版社，2004.

上接第 25 页

[7] Tyler Moore, Richard Clayton. The Impact of Incentives on Notice and Take-down[C]. In Proc. of the 7th Workshop on the Economics of Information Security, New Hampshire USA, June 25-28 2007: 1-24.

[8] Christian Ludl, Sean McAllister, Engin Kirda,et al.. On the Effectiveness of Techniques to Detect Phishing Sites[C]. In Proc. of the 4th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment, Lucerne Switzerland, July 12-13 2007:20-39.

[9] Steve Sheng, Brad Wardman, Gary Warner, et al.. An Empirical Analysis of Phishing Blacklists[C]. In Proc. of the sixth Conference on Email and Anti-Spam, California USA, July 16-17 2009.

[10] Pawan Prakash, Manish Kumar, Ramana Rao Kompella, et al.. PhishNet: Predictive Blacklisting to Detect Phishing Attacks[C]. In Proc. of the IEEE INFOCOM, San Diego Canada, March 14-19 2010:1-5.

[11] Ye Cao, Weili Han, Yueran Le. Anti-phishing Based on Automated Individual White-List[C]. In Proc. of the DIM'08, Virginia, USA, Oct. 31, 2008:51-59.

[12] Sujata Garera, Niels Provos, Monica Chew, et al.. A Framework for Detection and Measurement of Phishing Attacks[C]. In Proc. of the WORM'07, Virginia USA, Nov. 2, 2007:1-8.

[13] Colin Whittaker, Brian Ryner, Marria Nazif. Large-Scale Automatic Classification of Phishing Pages [C]. In Proc. of 17th Annual Network and Distributed System, California USA, Feb. 28-March 3 2010.

[14] Justin Ma, Lawrence K. Saul, Stefan Savage, et al. Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs[C]. In Proc. of the KDD'09, Paris France, June 28-July 1, 2009:1245-1254.

[15] Justin Ma, Lawrence K. Saul, Stefan Savage, et al. Identifying Suspicious URLs: an Application of Large-scale Online Learning[C]. InProc. of the 26th International Conference on Machine Learning, Montreal Canada, June, 14-18, 2009.

[16] Kurt Thomas, Chris Grier, Justin Ma, et al. Design and Evaluation of a Real-Time URL Spam Filtering Service[C]. In Proc. of the IEEE Symposium on Security and Privacy, California USA, May, 22-25, 2011.

[17] Aaron Blum, Brad Wardman, Thamar Solorio, et al. Lexical Feature based Phishing URL Detection using online Learning[C]. In: Proc. of the AISec'10, Chicago USA, Oct. 8 2010:54-60.

[18] Yue Zhang, Jason Hong, Lorrie Cranor. CANTINA: A Content-Based Approach to Detecting Phishing Web Sites[C]. In Proc. of WWW2007, Alberta Canada, May 8-12, 2007: 639-648.

[19] Anthony Y. Fu Liu Wenyin, Xiaotie Deng, et al.. Detecting Phishing Web Pages with Visual Similarity Assessment Based on Earth Mover's Distance (EMD)[J]. IEEE Transaction on Dependable and Secure Computer, 2007, 3(04):301-311.

[20] Chih-Chung Chang, Chih-Jen Lin. LIBSVM: a Library for Support Vector Machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 3(02):1-27.