文章编号：1000-8608(2013)06-0903-05

基于敏感特征的网络钓鱼网站检测方法

宋明秋*， 曹晓芸

（大连理工大学 管理科学与工程学院，辽宁 大连 116024）

摘要： 网络钓鱼(phishing)是一种在线欺诈行为，普遍存在于电子商务和电子金融中. 将黑白名单方法和异常特征检测方法相结合，针对网络钓鱼网站URL异常和页面身份异常特点提出基于敏感特征的网络钓鱼网站检测方法——PhishDetector. 使用黑白名单技术对URL进行拦截，对于名单中不存在的URL，提取其敏感特征，然后使用线性分类器判断该网站是否为网络钓鱼网站. 实验结果表明，基于敏感特征的网络钓鱼网站检测方法，提高了网络钓鱼网站检测的正确率，显著降低了误判率.

关键词： 网络钓鱼；敏感特征；线性分类器

中图分类号： TP393.08 **文献标志码：** A

0 引 言

巨大的经济利益使得网络钓鱼攻击的数量与日俱增，其危害范围也不断扩大. 根据中国反钓鱼联盟的报告，截至2012年2月，联盟认定并处理了网络钓鱼网站80 076个，相比2011年处理掉的43 842个，增长了83%[1]. 根据Gartner的调查报告，截至2007年8月的一年中，有360万名成人误入钓鱼陷阱，由此造成的直接经济损失超过30亿美元，间接经济损失可能会更高[2].

目前已出现Google Safe Browsing、Microsoft Phishing Filter等10余种反钓鱼工具，主要采用了黑白名单[3]和异常特征检测[4]两种方法. 然而这两种方法都具有一定的局限性：黑白名单技术不能及时更新最新的网络钓鱼网站的黑名单，表现出一定的滞后性；异常特征检测技术所采用的特征大多取自已出现的攻击，比较具体明确，容易被攻击者绕过，造成负误判率比较高.

本文针对上述两种方法存在的问题，将黑白名单和异常特征检测方法相结合，对异常特征的选取以及URL拦截部分分别进行改进，提出基于敏感特征的PhishDetector检测方法，以有效降低误判率.

1 基于敏感特征的网络钓鱼检测

网络钓鱼网站与合法网站相比，数量上差距极大且存活期短，因而单纯的只检测Web页面特征异常，无法检测出包含新特征的钓鱼网站，容易产生误判. 网络钓鱼的本质是利用用户的心理弱点哄骗用户给出机密信息，而钓鱼网址本身并不包含网络钓鱼的这一决定性特征，因而单纯采用黑白名单技术或者只检测URL异常无法完全确定其是否为钓鱼网站.

基于此，本文将Web页面特征异常和URL异常结合起来，分别提取URL异常和Web页面内容异常的敏感特征，提出PhishDetector检测方法，方法流程如图1所示. 该方法基本分为3个阶段：首先使用黑白名单技术提高检测速度，减少误判率；其次对于不属于黑白名单的URL地址页面提取敏感特征；最后使用线性分类器判断所检测的网站是否为网络钓鱼网站.

对于URL拦截部分的改进主要是引入了布隆过滤器的概念，将判断请求访问的URL是否在黑白名单之中转化为检索一个元素是否在一个(URL地址)集合之中.

收稿日期： 2012-10-15； **修回日期：** 2013-09-06.
基金项目： 国家自然科学基金资助项目(71171028).
作者简介： 宋明秋*(1967-)，女，博士，副教授，E-mail：songmq@dlut.edu.cn.

对于异常特征的选择,本文与已有基于 URL 结构特征、词汇特征[5]等特征检测方法不同,提出基于 URL 及 Web 页面敏感特征的网络钓鱼网站检测方法,具体方法描述如下.

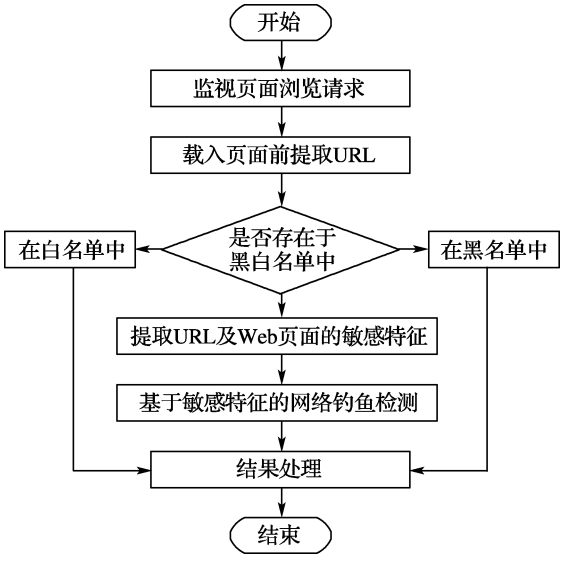

图 1 网络钓鱼攻击检测流程

Fig. 1 The flow-chart of network phishing detection

1.1 URL 敏感特征提取

URL 网址正常情况下只能使用英文字母、阿拉伯数字和某些标点符号,不能使用其他文字和符号,如:http://www. jb51. net/article/16271. htm 为一个正常的 URL,如果出现其他符号(如汉字)或特殊标识(如@等)则为 URL 异常. 网络钓鱼网站 URL 中的敏感特征检测主要包括以下 5 个方面:

(1)域名中是否包含 IP 地址[6]. 有的网络钓鱼网站会采用 IP 地址代替部分域名.

(2)URL 中“.”的个数. 检查 URL 中是否包含过多的“.”,网络钓鱼网站中“.”的个数较合法网站要多(通常≥5).

(3)URL 端口. 部分网络钓鱼攻击者采用非 80 端口代替传统的 http 协议的 80 端口,目的在于逃脱反钓鱼工具的检测.

(4)URL 中异常字符. 检测 URL 中是否包含“@”或“—”等异常字符.

(5)域名注册年龄. 为避免钓鱼网站地址出现在黑名单之内,网络钓鱼攻击通常发生在钓鱼网站注册很短的时间内,比如几天之内. 一般认为域名注册时间超过 12 个月的网站为合法网站.

1.2 Web 页面敏感特征提取

网络钓鱼网站虽然具有一定的共性,但实际上存在着差异[7]. 从 Web 页面的 DOM 文档对象模型和 HTTP 协议分析的角度,网络钓鱼网站的页面存在几种异常:链接 URL 异常、表单异常、资源引用异常、域名信息异常以及 Cookie 异常.

由于网络钓鱼攻击是由攻击者通过伪造某一个合法网站的身份引起的[8],在提取 Web 页面敏感特征前,需要先提取 Web 页面所伪造的身份特征. 提取 Web 页面身份特征的基本思想是:依据词出现的频率来确定关键词[9]. 本文使用 Pan 的身份提取算法[10],从 Web 网页相关的 DOM 对象和属性中提取 5 个身份特征:ICP 证号、网页域名与 URL 身份的一致性、链接对象、资源引用异常和 Form 表单异常.

(1)Web 网站的 ICP 证号

ICP 许可证是指由各地通信管理部门核发的《中华人民共和国电信与信息服务业务经营许可证》,可用来唯一地标识网站身份[10]. 图 2 为 DELL 中国公司某网站页面,在页面底部圈住的部分是该网站的 ICP 许可证号.

图 2 DELL 中国页面

Fig. 2 The homepage of DELL China

(2)网页域名与 URL 身份的一致性

正常合法网页的 URL 身份就是该网页的域名,网络钓鱼网页的 URL 身份通常都是攻击者要模仿的合法网站的域名.

(3)链接对象

本文所说的链接对象包括 3 类. ①空链接对

象：Web 页面中链接对象指向为空的链接；②指向真实站点的链接对象：指向的域名与页面所在的域名不一致的链接对象；③指向本域的链接对象：指向的域名和该页面所在的域名一致的链接对象[10].

(4)资源引用异常

指网络钓鱼网站页面中存在的图片来源与页面所表现出的所有者的域不同.

(5)Form 表单异常

合法网页中，Form 表单将用户输入数据提交到本域内服务器端，其行为与该网站页面表现出的身份一致；而网络钓鱼网站则常常将用户提交的数据提交到钓鱼服务器，用户信息被攻击者获得. 此外，网络钓鱼网站页面有可能会出现空的 Form 表单，因此，Form 表单可用于检测网页异常.

2　算法构建

2.1　敏感特征函数

将 URL 和 Web 页面的敏感特征组成敏感特征向量 $\boldsymbol{F}$，$\boldsymbol{F}=(F_1\ \ F_2\ \ F_3\ \ F_4\ \ F_5\ \ F_6\ \ F_7\ \ F_8\ \ F_9\ \ F_{10})$. 其中 F_1 为 URL 中是否含有 IP 地址；F_2 为 URL 中“.”的数量；F_3 为 URL 端口；F_4 为 URL 中的异常字符；F_5 为 URL 的域名注册年龄；F_6 为网页的 ICP 证号；F_7 为网页域名与 URL 身份的一致性；F_8 为链接对象；F_9 为资源引用；F_{10} 为 Form 表单.

特征 F_1 到 F_5 为网络钓鱼网站的 URL 敏感特征，F_6 到 F_{10} 则为网络钓鱼网站页面的敏感特征. 把这些特征用函数形式来表示：

$$F_1=\begin{cases}1; & \text{URL 中含有 IP 地址}\\ -1; & \text{其他}\end{cases}\tag{1}$$

$$F_2=\begin{cases}1; & \text{URL 中“.”的数量}\geqslant 5\\ -1; & \text{其他}\end{cases}\tag{2}$$

$$F_3=\begin{cases}1; & \text{URL 端口非 80}\\ -1; & \text{URL 端口为 80}\end{cases}\tag{3}$$

$$F_4=\begin{cases}1; & \text{URL 中含有“@”等异常字符}\\ -1; & \text{其他}\end{cases}\tag{4}$$

$$F_5=\begin{cases}1; & \text{其他}\\ -1; & \text{URL 的域名注册年龄超过 12 个月}\end{cases}\tag{5}$$

$$F_6=\begin{cases}1; & \text{给定网页的 ICP 与某一合法网}\\ & \text{站的 ICP 相同，但域名不同}\\ -1; & \text{其他}\end{cases}\tag{6}$$

$$F_7=\begin{cases}1; & \text{其他}\\ -1; & \text{网页的 URL 身份与网页的}\\ & \text{域名一致}\end{cases}\tag{7}$$

$$F_8=\begin{cases}(N_n+N_r)/N_a; & N_n+N_r\geqslant N_l>0\\ 0; & N_a=0\\ -N_l/N_a; & N_l\geqslant N_n+N_r>0\end{cases}\tag{8}$$

式中：N_a 为网页中链接对象的总数，N_n 为网页中空链接的链接对象个数，N_r 为指向真实站点的链接对象个数，N_l 为指向本域的链接对象的个数. 如果 $N_n+N_r\geqslant N_l>0$，该网站页面的异常链接所占比重大，则 $F_8=(N_n+N_r)/N_a$，其值越趋近于 1，网页异常程度越大；$N_a=0$，该网站页面中没有链接，因此 $F_8=0$；如果 $N_l\geqslant N_n+N_r>0$，表明指向本域的链接对象所占的比重大，$F_8=-N_l/N_a$，其值越趋近于 −1，该网站页面越正常.

$$F_9=\begin{cases}P_f/P_a; & P_f>P_l>0\\ 0; & P_a=0\\ -P_l/P_a; & P_l\geqslant P_f\geqslant 0\end{cases}\tag{9}$$

式中：P_a 为网站页面中的图片总数；P_f 为来自外域的图片个数；P_l 为来自本域的图片个数. 如果 $P_f>P_l>0$，该网页来自外域的图片所占比例大，则 $F_9=P_f/P_a$，其值越趋近于 1，异常程度越大；$P_a=0$，该网站页面中没有图片，因此 $F_9=0$；如果 $P_l\geqslant P_f\geqslant 0$，表明来自本域的图片所占的比重大，$F_9=-P_l/P_a$，其值越趋近于 −1，该网站页面越正常.

$$F_{10}=\begin{cases}1; & \text{出现空的数据提交对象}\\ & \text{或者身份不一致}\\ -1; & \text{其他}\end{cases}\tag{10}$$

$F_{10}=1$ 表示页面 Form 表单的数据提交对象存在异常，可能是网络钓鱼页面；否则，表示正常.

2.2 线性分类函数

根据上述敏感特征函数，判断一个网站是否为网络钓鱼网站时，使用线性分类器进行处理：

$$S = f(\sum F_i \times w_i);\ 0 < i \leqslant 10 \quad (11)$$

其中

$$f(x) = \begin{cases} 1; & x > 0 \\ -1; & x \leqslant 0 \end{cases}$$

$F_i(1 \leqslant i \leqslant 10)$ 表示网络钓鱼攻击敏感特征的取值. $f(x) = 1$ 时，判断该网站为网络钓鱼攻击网站；$f(x) = -1$ 则表示页面正常，该网站为非网络钓鱼攻击网站. $w_i(1 \leqslant i \leqslant 10)$ 为 10 个敏感特征分量(F_i) 的权值，其相应的计算公式如下：

$$w_i = \frac{e_i}{\sum_{i=1}^{10} e_i} \quad (12)$$

$$e_i = T_{F_i} - F_{F_i} \quad (13)$$

式中：T_{F_i} 和 F_{F_i} 分别为单独使用敏感特征 F_i 检测页面时的正确率和误判率.

3 实验结果及其分析

为了确定每个敏感特征的最佳权值，从网络钓鱼网站数据库 Phishtank[11] 中随机选取了 50 个钓鱼网站和 50 个合法网站进行实验，实验结果如表 1 所示.

表 1 敏感特征提取结果

Tab. 1 Results of sensitive characteristics extraction

%

敏感特征项	T_{F_i}	F_{F_i}	e_i	w_i
IP 地址	42	5	37	8.0
URL 中“.”数	49	5	44	9.5
URL 端口	89	12	77	16.7
异常字符	7	2	5	1.1
域名注册年龄	85	33	52	11.3
ICP 证号	98	11	87	18.9
域名与 URL 身份的一致性	72	15	57	12.4
链接对象	34	19	15	3.3
资源引用异常	31	7	24	5.2
Form 表单异常	96	33	63	13.7

由表 1 可知，ICP 证号和 URL 端口的权值最大，分别为 18.9%和 16.7%；加上 Form 表单、域名与 URL 身份的一致性，以及域名注册年龄，共有 73%的权值可以识别网络钓鱼网站. 而链接对象和资源引用异常的权值比较小，分别为 3.3%和 5.2%.

为了测试检测方法的有效性，选取网络钓鱼检测工具 SpoofGuard 和 Google Safe Browsing，与 PhishDetector 检测方法进行比较. 测试数据同样是从 Phishtank 中随机抽取的 50 个网络钓鱼网站和 50 个合法网站，选取正确率与误判率作为评价指标，实验结果如图 3 所示.

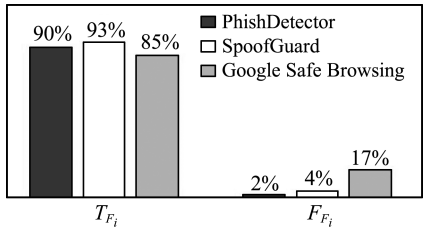

图 3 PhishDetector 和 SpoofGuard、Google 检测工具的比较结果

Fig. 3 Comparison of PhishDetector, SpoofGuard and Google Detection tools

通过实验结果可以发现，SpoofGuard 正确率最高，可达到 93%，而 Google Safe Browsing 误判率最高，达到了 17%；改进后的 PhishDetector 检测算法正确率达到了 90%的水平，而误判率显著降低到 2%，说明改进后的方法对于网络钓鱼网站的识别具有很好的效果.

4 结 语

本文在传统的基于黑白名单和异常特征检测技术方法基础之上，提出网络钓鱼网站的敏感特征的概念及定量描述方法，并提出基于敏感特征的网络钓鱼网站检测算法 PhishDetector. 算法结合网络钓鱼网站的 URL 异常及 Web 页面的身份异常，提取敏感特征，使用线性分类器对可疑网站进行分类. 实验结果表明，PhishDetector 算法对网络钓鱼网站的识别较之 SpoofGuard 误判率要低 15%，相比 Google Safe Browsing 的检测正确率要高，证明了算法的有效性. 对于敏感特征的自动识别将是未来研究的主要问题.

参考文献：

[1] 中国反钓鱼网站联盟. 2012 年 3 月钓鱼网站处理

简报[R]. 北京:APAC, 2012.
Anti-phishing Alliance of China(APAC). Phishing websites handling bulletin [R]. Beijing: APAC, 2012. (in Chinese)

[2] Stamford C. Gartner survey shows phishing attacks escalated in 2007; more than $3 billion lost to these attacks [EB/OL]. (2007-12-17). http://www. gartner. com/newsroom/id/565125.

[3] 郭敏哲. 基于浏览器的网络钓鱼检测机制的研究与实现[D]. 北京:北京林业大学, 2008.
GUO Min-zhe. The study and implementation of browser-based phishing detection [D]. Beijing: Beijing Forestry University, 2008. (in Chinese)

[4] Microsoft Corporation. Microsoft phishing filter: a new approach to building trust in E-commerce content [R]. Redmond: Microsoft Corporation, 2008.

[5] 黄华军,钱　亮,王耀钧. 基于异常特征的钓鱼网站URL检测技术[J]. 网络信息安全, 2012(1):23-25.
HUANG Hua-jun, QIAN Liang, WANG Yao-jun. Detection of phishing URL based on abnormal feature [J]. **Netinfo Security**, 2012(1):23-25. (in Chinese)

[6] Thelwall M, Wilkinsona D. Generic lexical URL segmentation framework for counting links, colinks or URLs [J]. **Library & Information Science Research**, 2008, **30**(2):94-101.

[7] 石　静. 网络内容过滤系统的设计与实现[D]. 上海:复旦大学, 2005.
SHI Jing. Design and implementation of network content filtering system [D]. Shanghai: Fudan University, 2005. (in Chinese)

[8] Prakash P, Kumar M, Kompella R R, *et al*. Phishnet: predictive blacklisting to detect phishing attacks [C] // **INFOCOM'10 Proceedings of the 29th Conference on Information Communications**. Piscataway: IEEE Press, 2010.

[9] ZHANG Yue, Hong J, Cranor L F. Cantina: A content-based approach to detecting phishing web sites [C] // **16th International World Wide Web Conference, WWW2007**. New York: Association for Computing Machinery, 2007:639-648.

[10] PAN Ying, DING Xu-hua. Anomaly based web phishing page detection [C] // **Proceedings - Annual Computer Security Applications Conference, ACSAC**. Los Alamitos: IEEE Computer Society, 2006:381-390.

[11] Phish Tank. Phishtank [DB/OL]. (2006-11-20). http://www. phishtank. com/.

Phishing detection method based on sensitive characteristics of phishing webpage

SONG　Ming-qiu*,　CAO　Xiao-yun

(School of Management Science and Engineering, Dalian University of Technology, Dalian 116024, China)

Abstract: Phishing is a kind of online fraud which is widespread in the electronic commerce and electronic banking. A method of phishing detection based on sensitive characteristics — PhishDetector is proposed. Firstly, the black/white list technology is used to intercept URL, and then the sensitive features of those URL not existing in the lists are extracted, and lastly a phishing webpage is distinguished with a linear classifier. The experimental results show that this method is better both in accuracy and false positive rate.

Key words: phishing; sensitive characteristics; linear classifier