

# 利用图像识别技术过滤海量可疑钓鱼网站

周诚诚<sup>1,2</sup>, 张代远<sup>1,2,3</sup>

- (1. 南京邮电大学 计算机学院, 江苏 南京 210003;  
2. 江苏省无线传感网高技术研究重点实验室, 江苏 南京 210003;  
3 南京邮电大学 计算机技术研究所, 江苏 南京 210003)

**摘要:**网络钓鱼攻击(phishing, 又称钓鱼攻击、网络钓鱼)作为一种主要基于互联网传播和实施的新兴攻击、诈骗的方式, 正呈逐年上升之势, 使广大用户和金融机构遭受到财产和经济损失。如何及时、有效地识别网络钓鱼相关的互联网风险, 控制钓鱼攻击可能带来的影响, 已经成为各金融机构当前亟待解决的问题。因此, 各大银行、证券公司以及安全公司纷纷推出自己的反钓鱼监控服务, 目前的反钓鱼技术普遍采取利用爬虫主动进行大范围互联网仿冒站点的搜索, 爬取大量可疑钓鱼网站, 并逐一一对可疑钓鱼网站进行检测, 判断其是否为钓鱼网站。面对海量可疑网站, 如何高效快速地检测出可疑钓鱼网站又成为一个难题。文中介绍了一种基于图像识别技术的网站徽标(LOGO)检测的新思路, 用于过滤海量的可疑钓鱼网站, 加快钓鱼网站的检测效率。

**关键词:**图像识别; 钓鱼网站; 可疑钓鱼网站; phishing 反钓鱼技术; LOGO 检测

中图分类号: TP309

文献标识码: A

文章编号: 1673-629X(2012)11-0246-04

## Using Image Recognition Technology to Filter Mass Suspicious Phishing Sites

ZHOU Cheng-cheng<sup>1,2</sup>, ZHANG Dai-yuan<sup>1,2,3</sup>

- (1. College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;  
2. Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing 210003, China;  
3. Institute of Computer Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

**Abstract:** Phishing attacks (phishing, also known as phishing attacks), as an emerging attacking and frauding way primarily based on internet for dissemination and implementation, is burgeoning increasingly year by year, so that customers and financial institutions have suffered property and economic losses. How to identify the internet risks as to phishing and control the possible impact brought by phishing attacks timely and effectively, has become the current problems to be solved by various financial institutions. Therefore, the major banks, security companies and the major safeco have successively launched their own anti-phishing monitoring services. Present anti-phishing technology generally take advantage of spiders to conduct a wide-range search of fake internet sites initiatively, to crawl a large number of dubious phishing sites and to detect them one by one to determine whether it is a phishing site. However, it has become another problem how to detect phishing sites fast and efficiently facing massive dubious websites. It introduces a new idea, website logo detection in the light of image recognition technology, filtering massive suspicious phishing sites and accelerating the efficiency of detecting phishing sites.

**Key words:** image recognition; phishing sites; dubious phishing sites; anti-phishing technology; LOGO detection

## 0 引言

近年来, 互联网在中国得到蓬勃的发展, 但越来越严重的安全问题引起了人们的担忧, 其中, 网络钓鱼攻击成为网络攻击的代表。网络钓鱼是一种企图通过伪

造真实的网页来欺骗使用者, 从而获取使用者的银行账号、密码等个人敏感信息的犯罪诈骗过程。2011年第一季度, 针对银行发起的钓鱼攻击事件大规模爆发, 短时间内新增银行类钓鱼网站近千个, 至少数十万用户访问过此类网站, 造成的经济损失达数亿元<sup>[1]</sup>。

收稿日期: 2012-03-08; 修回日期: 2012-06-11

基金项目: 江苏高校优势学科建设工程资助项目(yx002001)

作者简介: 周诚诚(1987-), 男, 硕士研究生, 研究方向为智能计算技术与应用; 张代远, 教授, 硕士生导师, 研究方向为智能计算理论、方法与应用, 计算机体系结构, 计算机在通信中的应用。

## 1 反钓鱼技术分析

### 1.1 目前钓鱼网站的检测方法和流程图

随着互联网的发展, 出现了大量域名极为相似的

钓鱼网站(如 [www.taobao.com](http://www.taobao.com) 等),这类钓鱼网站与正规网站域名相似,具有很大的欺骗性。

目前采用的反钓鱼技术还不能完全及时有效地检测出互联网中所有的钓鱼网站,但针对银行、证券等金融机构的正规网站,所采取的一般检测思路是:

1、将它们的域名全部收集起来,建立一个信誉库(相当于白名单)加以保护。

2、通过一种搜索策略,利用爬虫技术不断从互联网中爬取可疑钓鱼网页。

3、通过各种检测方法找到和信誉库中网页“相似的”页面,如果相似度超过一定的阈值,就可以确定该网站为钓鱼网站。

4、将钓鱼网站加入到黑名单中去,并上报给有关组织<sup>[2]</sup>。

在步骤 3 中对这种“相似页面”的钓鱼检测的思路主要有以下几种:

1、判定网页结构、内容的相似度。

2、将网页看作图片,比较两个图片的相似度<sup>[3]</sup>。

3、将上面两种方法进行一定的融合。

图 1 给出了钓鱼网站检测的流程<sup>[4,5]</sup>。

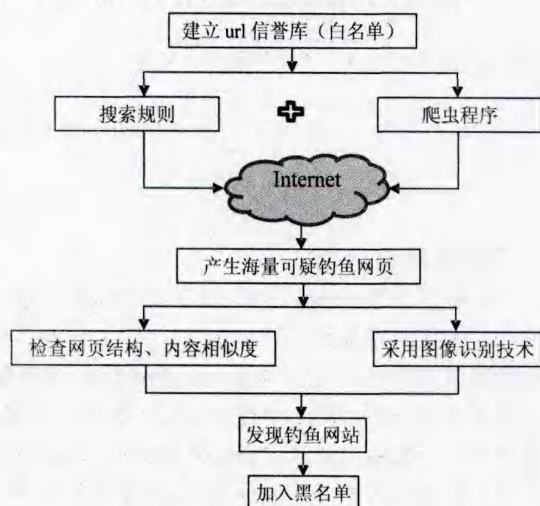


图 1 钓鱼网站检测流程

## 1.2 目前反钓鱼技术的缺陷及改进方法

从图 1 可以看到,对海量可疑钓鱼网站的检测在这个过程中,不管采用哪种判别方法,都要逐一比较可疑站点和受保护站点之间的相似度,对海量可疑钓鱼网站这种逐一比较的方式难免会导致检测效率过低,不能有效及时响应钓鱼攻击,并且会导致机器负载增大。为了能有效及时地处理钓鱼攻击,有必要对海量的可疑钓鱼网站进行过滤,毕竟在整个互联网中,钓鱼网站的所占比例还是比较小的,爬虫程序爬取的海量钓鱼网站中只有一小部分才是真正的钓鱼网站,那么如何对海量可疑站点进行过滤呢?下面将要给出一种基于图像识别技术过滤海量钓鱼网站的方法。

攻击者要想钓鱼攻击成功,其制作的伪造页面一定要在视觉上和原始页面很像,这才能很大程度上提高钓鱼的成功率。同时也注意到大部分的被钓鱼的组织或者机构的官方网站都会包含相应的组织徽标 LOGO,由于伪造页面和原始页面十分相似,因此在很大程度上伪造的页面也会包含相应的 LOGO。

最近有关中国光大银行钓鱼事件,超级巡警安全中心对中国光大银行钓鱼网站进行了完整分析,从页面上分析来看,中国光大银行钓鱼网站和银行官方网站几乎一模一样,而且都含有官方的 LOGO,就连网址也很像(钓鱼网:<http://cebbrnk.com>,中国光大银行官网:<http://cebbank.com>),很难区分,虽然网址只差一个字母,但它们所指向的 IP 地址却各不相同,粗心的用户一不小心就上当了。

正是基于此点原因,如果能够利用图像识别技术,在海量的网页中迅速地筛选出包含有特定 LOGO 的网页,必将大大地提升发现钓鱼站点的效率。下面给出利用图像识别技术过滤海量可疑钓鱼网站的思路<sup>[6]</sup>:

1、在建立信誉库的时候,不仅要保存受保护站点的 url,还要保存站点的 LOGO 图片。

2、将爬虫爬取的海量可疑钓鱼网站全部变成图片。

3、利用图像识别方法从海量可疑钓鱼网站中识别出该网站是否含有信誉库中的 LOGO,删除不匹配的可疑钓鱼网站的 url 和图片。

4、将删除后的可疑钓鱼网站进行“相似页面”的检测,判断其是否为钓鱼网站。

从上面给出的思路中看出,过滤的关键技术就是第三步中如何利用图像识别方法从海量可疑钓鱼网站中识别出该网站是否含有 LOGO,下面给出具体的图像识别算法设计。

## 2 图像识别算法设计

图像识别是人工智能和机器学习<sup>[7]</sup>的一个重要且成熟的研究领域,在工业中有广泛的应用,比如人脸识别、车牌检测和验证码识别等领域<sup>[8]</sup>。

在一张网页截图中识别出 LOGO,可以看作是一种图像匹配问题,所谓图像匹配,就是指图像之间的比较,得到不同图像之间的相似度。图像匹配技术是数字图像处理领域的一项重要研究,文中的 LOGO 匹配采取的是一种基于灰度的模板匹配算法。

所谓基于灰度的模板匹配<sup>[9]</sup>,指的是首先将待匹配图像和源图像进行灰度化处理,然后通过模板匹配算法计算出相应的匹配系数(匹配系数的范围是大于 0 并且小于 1 的,匹配系数越接近 1,那么两个图像匹配相似度就越高),根据匹配系数得出匹配图像在

源图像中是否存在相似的区域,若存在相似的区域,在源图像中标识出待匹配图像;反之,则在源图像中不予标识。

模板匹配常用的一种计算方法是计算模板与源图像对应区域的误差平方和<sup>[10]</sup>。设  $f(x, y)$  为  $M \times N$  的源图像,  $t(j, k)$  为  $J \times K$  ( $J \leq M, K \leq N$ ) 的模板图像,其中  $M, N, J, K$  代表图像像素,  $x, y, j, k$  代表像素位置,则误差平方和测度定义为:

$$D(x, y) = \sum_{j=0}^{J-1} \sum_{k=0}^{K-1} [f(x+j, y+k) - t(j, k)]^2 \quad (1)$$

由(1)式展开可得:

$$D(x, y) = \sum_{j=0}^{J-1} \sum_{k=0}^{K-1} [f(x+j, y+k)]^2 - 2 \sum_{j=0}^{J-1} \sum_{k=0}^{K-1} t(j, k) \cdot f(x+j, y+k) + \sum_{j=0}^{J-1} \sum_{k=0}^{K-1} [t(j, k)]^2 \quad (2)$$

令:

$$DS(x, y) = \sum_{j=0}^{J-1} \sum_{k=0}^{K-1} [f(x+j, y+k)]^2 \quad (3)$$

$$DT(x, y) = \sum_{j=0}^{J-1} \sum_{k=0}^{K-1} [t(j, k)]^2 \quad (4)$$

$$DST(x, y) = 2 \sum_{j=0}^{J-1} \sum_{k=0}^{K-1} [t(j, k) \cdot f(x+j, y+k)] \quad (5)$$

$DS(x, y)$  与像素位置  $(x, y)$  有关,随着像素位置  $(x, y)$  的变化而变化,但会越来越缓慢。 $DST(x, y)$  也是随着像素位置  $(x, y)$  的变化而变化,当其取最大值时也就是模板  $t(j, k)$  和源图像中对应区域相匹配。 $DT(x, y)$  与图像像素位置  $(x, y)$  无关,只用一次计算便可。采用计算误差平方和测度可以减少计算量。

在(2)式中,若设  $DS(x, y)$  为常数,则图像匹配的参数只和  $DST(x, y)$  有关,当  $DST(x, y)$  取最大值时,便可认为模板与图像是匹配的。但假设  $DS(x, y)$  为常数,这种情况下会产生一定的误差,严重时将无法准确匹配,因此可用另一种方法也就是归一化互相关作为误差平方和测度<sup>[11]</sup>,其定义见(6)式:

$$R(x, y) = \frac{\sum_{j=0}^{J-1} \sum_{k=0}^{K-1} t(j, k) \cdot f(x+j, y+k)}{\sqrt{\sum_{j=0}^{J-1} \sum_{k=0}^{K-1} [f(x+j, y+k)]^2} \cdot \sqrt{\sum_{j=0}^{J-1} \sum_{k=0}^{K-1} [t(j, k)]^2}} \quad (6)$$

图2给出了模板匹配的示意图,在图中假设源图像  $f(x, y)$  和模板图像  $t(j, k)$  的原点都在左上角。对任何一个  $f(x, y)$  中的像素点  $(x, y)$ ,根据(6)式可以算出一个  $R(x, y)$ 。当  $x$  和  $y$  变化时,  $t(j, k)$  在源图像区域中移动并得出  $R(x, y)$  所有值。 $R(x, y)$  的最大值直接反映了模板图像和源图像匹配的最佳位置<sup>[12]</sup>,若从该位置开始在源图像中取出与模板大小相同的区域,便可找到匹配图像。

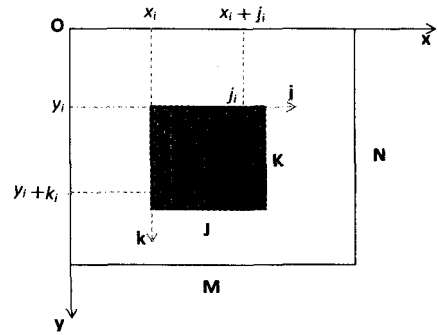


图2 模板匹配示意图

图3给出了模板匹配的流程图。

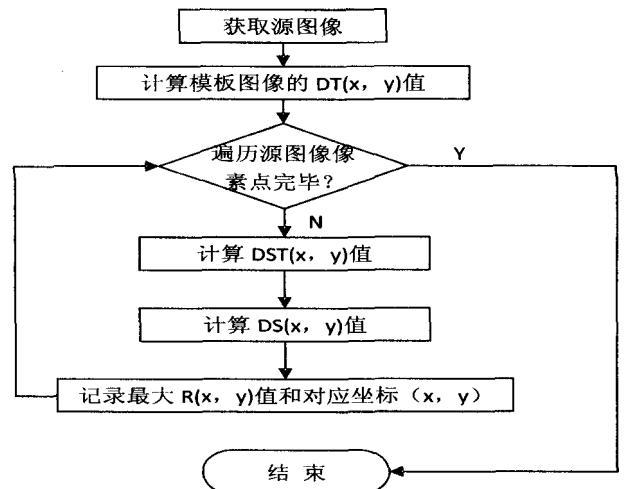


图3 模板匹配流程图

### 3 实验仿真与分析

本实验是在 Windows 7+Matlab 7.0 的环境下进行仿真的。采用的是 Matlab 的 GUI 设计,图4给出了匹配后的结果,其中左边的区域显示的是源图片,这里选取的是光大银行钓鱼网站中包含 LOGO 的部分截图,大小为  $366 \times 55$ ,右边的区域显示的是模板图像,选取的是光大银行官方网站 LOGO 截图,大小为  $81 \times 49$ ,按照基于灰度的模板匹配算法,匹配成功后,在光大银行的可疑钓鱼页面将 LOGO 用红色标记出,整个匹配所需要的时间约为 3.7 秒。

从图4的实验结果看出匹配所需时间还是一个令人接受的结果,但如果选取的源图像大小为  $600 \times 364$ ,模板图片选取的大小为  $360 \times 55$ ,进行上面的模板匹配实验,所花费的时间大概需要 123 秒,这个匹配所需时间比较长了,匹配效率有点低。

初步的实验结果证明了采用 LOGO 匹配方法的有效性,只要可疑钓鱼网站的页面中含有 LOGO 图片,就能被识别出来,不足之处就是在源图片比较大的情况下,匹配所需要的时间是比较久,如何改进匹配算法,提高匹配效率,是今后努力的方向。

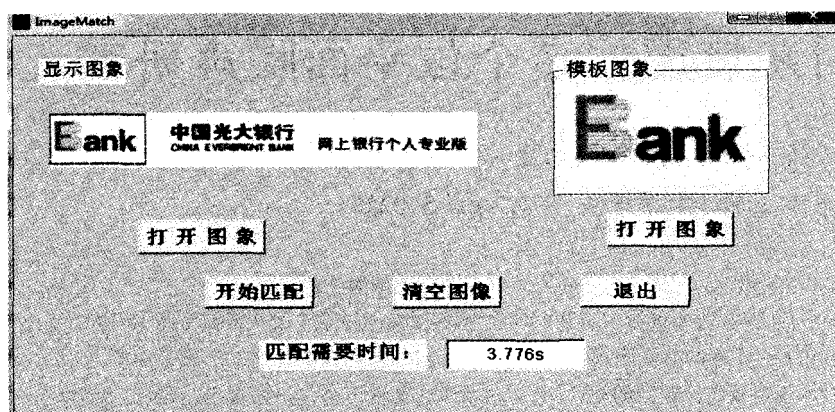


图 4 LOGO 匹配结果截图

## 4 结束语

文中介绍了钓鱼网络攻击以及图像识别技术在钓鱼检测过程中的应用,提出了一种通过基于网站徽标 LOGO 识别来过滤海量钓鱼页面的新思路,通过实验证明了思路的有效性。对于 LOGO 识别采取的基于灰度的模板匹配算法而言,时间复杂度较高,对于大型的图片,匹配的时间过长,如何改进算法,加快 LOGO 识别的效率,这是下一步亟待解决的问题。

### 参考文献:

- [1] 华为赛门铁克科技有限公司. 网络钓鱼专题报告[R]. 成都:华为赛门铁克科技有限公司,2011.
- [2] 李佟鸿,麦永浩. 网络钓鱼犯罪技术与对策研究[J]. 信息安全,2011(4):24-26.
- [3] 曹玖新,罗军舟,毛波. 基于图像处理的钓鱼网页检测方

法[P]. 中国, CN 200710130809, 2008-03-19.

- [4] 林世飞,杨勇,马松松,等. 一种钓鱼网站的检测方法及其装置[P]. 中国, CN 200910106559, 2009-09-16.
- [5] 张卫丰,周毓明,许蕾,等. 基于匈牙利匹配算法的钓鱼网页检测方法[J]. 计算机学报,2010(10):1963-1975.
- [6] 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]. 信息安全,2012(1):77-83.
- [7] Valiant L G. A Theory of the Learnable[J]. Communications of the ACM,1984,27(11):1134-1142.
- [8] Pan Ying, Ding Xuhua. Anomaly Based Web Phishing Page Detection[C]//Proc. of the 22nd Annual Computer Security Applications Conference. New Orleans, LA, USA; [s. n.], 2006:381-392.
- [9] 李强,张钹. 一种基于图像灰度的快速匹配算法[J]. 软件学报,2006,16(2):216-222.
- [10] Zen Chen. A Zernike Moment Phase-based Descriptor for Local Image Representation and Matching[J]. Image Processing,2010,19(1):205-219.
- [11] Freund Y. Boosting a Weak Learning Algorithm by Majority Information and Computation[J]. Information and Computation,1995,121(2):197-227.
- [12] Fisher Y. Fractal image compression with quadrees[M]//Fractal Image Compression: Theory and Application. New York:Springer-Verlag,1995:55-77.

(上接第 245 页)

辑更清晰,更便于修改、维护。

下一步工作中,将主要研究如何采用 Thin-Fat 机制以实现可信芯片的内核精简性问题。

### 参考文献:

- [1] 工业和信息化部电信研究院. 移动互联网白皮书[EB/OL]. 2011-05. HTTP://labs.chinamobile.com/report/66210.
- [2] 罗军舟,吴文甲,杨明. 移动互联网:终端、网络与服务[J]. 计算机学报,2011,34(11):2029-2051.
- [3] 程桂花,齐学梅,罗永龙. AES 算法中模逆运算电路设计与实现[J]. 小型微型计算机系统,2011,32(6):1240-1244.
- [4] 程桂花,齐学梅,罗永龙. AES 算法中多项式模运算及其性能分析[J]. 计算机技术与发展,2010,20(9):115-118.
- [5] Wu C H, Wu C M, Shieh M D. High-speed low-complexity systolic designs of novel iterative division algorithms in GF(2m)[J]. IEEE Trans on Computers,2004,53(3):375-380.

- [6] Tawalbeh L A, Tenca A F, Park S. A dual-field modular division algorithm and architecture for application specific hardware[J]. Systems and Computers,2004,1(7):483-487.
- [7] Guo J H, Wang C L. Systolic array implementation of Euclid's algorithm for inversion and division in GF(2m)[J]. IEEE Trans on Computers,1998,47(10):1161-1167.
- [8] 李雪梅,欧海文,路而红,等. 基于 FPGA 快速 AES 算法 IP 核的设计与实现[J]. 计算机工程与应用,2006(24):84-86.
- [9] 张晓丰,樊启华,程红斌. 密码算法研究[J]. 计算机技术与发展,2006,16(2):179-184.
- [10] 郑东,王友仁,张岩. AES 中字节代换和列混合的硬件可逆设计[J]. 计算机技术与发展,2009,19(7):191-194.
- [11] 黄智颖,冯新喜,张焕国. 高级加密标准 AES 及其实现技巧[J]. 计算机工程与应用,2002(9):112-115.
- [12] 刘小平,何云斌,量怀国. 基于 VerilogHDL 的有限状态机设计与描述[J]. 计算机工程与设计,2008,29(4):958-960.