

基于页面敏感特征的金融类钓鱼网页检测方法

胡向东¹, 刘可¹, 张峰², 林家富¹, 付俊², 郭智慧²

(1. 重庆邮电大学自动化学院, 重庆 400065;

2. 中国移动通信有限公司研究院, 北京 100033)

摘要: 提出一种基于页面敏感特征的金融类钓鱼网页检测方法, 通过获取网页超文本标记语言特定标签中的文本信息, 利用适合中文的多模式匹配算法 (AC_SC, AC suitable for Chinese) 匹配出敏感文本条数, 计算出敏感文本特征值; 定位截取网页的 logo 图像, 采用 PCA-SIFT 算法提取图像特征, 并与预先建立的网页 logo 图像库进行匹配, 计算出 logo 图像相似度; 基于文本特征值和图像相似度实现对金融类钓鱼网页的判定。实验结果表明, 该方法具有很强的针对性和时效性, 并能取得不低于 97% 的召回率。

关键词: 金融网页; 敏感特征; 文本特征值; 图像相似度; 钓鱼检测

中图分类号: TP393.08

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2017.00122

Financial phishing detection method based on sensitive characteristics of webpage

HU Xiang-dong¹, LIU Ke¹, ZHANG Feng², LIN Jia-fu¹, FU Jun², GUO Zhi-hui²

(1. School of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China;

2. Research Institute of China Mobile, Beijing 100033, China)

Abstract: A financial phishing detection method based on sensitive characteristics of webpage was proposed, which acquired sensitive text information of specific hypertext markup language tags and computes sensitive text eigenvalue. The method matches number of sensitive text using multiple pattern matching algorithm AC_SC (AC suitable for Chinese). Then, the method locates and cuts logo image of webpage, and utilizes PCA-SIFT algorithm to extract image features and match features with library of webpage logo which was established beforehand. Meanwhile, it calculates similarity of two logo image. Finally, the decision can be concluded by the text eigenvalue and image similarity. It shows that the method is better in pertinence and timeliness according to experiment, and achieves no less than 97% detection accuracy.

Key words: financial webpage, sensitive characteristic, text eigenvalue, image similarity, phishing detection

1 引言

随着互联网尤其是移动互联网的快速发展, 基于 Web 的应用深入各行各业, 为人们的工作、

生活带来了极大便利, 但用户个人信息被广泛收集, 面临严重的信息安全威胁, 如通过网络钓鱼骗取个人敏感信息进而窃取用户的钱财。根据 2016 年中国反钓鱼网站联盟 1~10 月的统计数据,

收稿日期: 2016-10-23; 修回日期: 2016-12-24。通信作者: 胡向东, huxd@cqut.edu.cn

基金项目: 教育部—中国移动联合研究基金资助项目 (No.MCM20150202); 重庆市教委科研基金资助项目 (No.KJ1602201)

Foundation Items: The Joint Research Foundation of the Ministry of Education of the People's Republic of China and China Mobile (No.MCM20150202), The Science and Technology Project Affiliated to Chongqing Education Commission (No.KJ1602201)

钓鱼网站涉及的行业前两位分别是支付交易类和金融证券类,其中仿冒中国工商银行、中国建设银行等金融机构的钓鱼网站数量一直处于前列,网络钓鱼存在攻击集中化的趋势^[1];亿盾互联2016年10月的钓鱼网站侦测报告也明确指出,银行是网络钓鱼的重灾区^[2]。

近年来,国内外专家、学者相继提出了一系列钓鱼网页检测方法,包括基于黑名单过滤方法、基于页面特征的启发式规则检测方法、基于页面视觉相似的检测方法等^[3]。文献[4]提出了一种黑名单改进方法 PhishNet,该方法利用黑名单中的 URL 作为先验知识,通过分解 URL 并计算相似性来检测钓鱼网页。文献[5]提出了一个基于页面内容检测的 CANTINA 方法,采用 TF-IDF 算法提取页面的 5 个关键词,通过检索关键词查看该网页域名是否在谷歌搜索返回结果的前 30 条以内,进行钓鱼网页检测,检测准确率约为 95%。该方法实施的根据是钓鱼网站存活时间短,导致其在搜索引擎中的排名很低甚至没被收录。CANTINA+^[6]是一种特征冗余的机器学习框架,是 CANTINA 的升级版。它通过提取 URL 特征、域名排名特征和页面内容特征等 15 个特征进行机器学习,特征考虑比较全面,提高了钓鱼网站检测准确率。文献[7]提取网页的文本特征、图像特征以及网页整体特征,利用匈牙利匹配算法计算出网页之间的相似性,具有较高的准确率和召回率。这些研究成果从不同角度、不同程度上提高了钓鱼网页识别的准确率。Sheng 等^[8]的研究指出 47%~83% 的钓鱼 URL 是在钓鱼事件发生 12 h 之后才被列入黑名单,有 63% 的钓鱼攻击在 2 h 内就已经结束。因此,基于黑名单过滤的方法无法应对最新出现的钓鱼网站。目前,金融类钓鱼网页通过精心制作,已无明显的 URL 特征,页面中被掺杂大量无意义的文本或字符,同时在关键的 HTML 标签中使用图像代替文本进行页面呈现,导致基于页面特征的启发式规则检测方法和基于页面视觉相似的检测方法无法提取到有用的特征,从而增加被正确检测识别的难度。

本文针对现有检测方法的不足以及金融类钓鱼网页规避检测的方法,对 PhishTank 公布的大量最新金融类钓鱼网站样本进行分析,将页面敏

感文本特征和 logo 图像特征相结合,提出了一种基于页面敏感特征的金融类钓鱼网页检测方法,有效地提高了检测准确率。

2 页面敏感特征选择

2.1 敏感文本特征

金融类网页中含有较多其独有的文本信息(如转账汇款、投资理财、网上银行等),称为敏感文本。钓鱼网页为了尽可能以假乱真,也会呈现出类似的敏感文本。通过分析大量的金融类网页以及钓鱼网页的源代码,发现敏感文本主要出现在页面 HTML 的 title、a、h、span 等特定标签中,可以对这些标签的文本信息进行提取,作为判定金融类钓鱼网页的依据。图 1 是手机版中国建设银行钓鱼网页,用方框圈出的文本为 a、h、span 等标签的文本。可以根据敏感文本出现次数的多少,来检测网页是否为钓鱼网页。考虑到大量非金融类网页 title 标签中也可能含有敏感文本,如金融机构的名称,所以不提取 title 标签中的文本。



图1 中国建设银行页面中特定标签的文本信息

敏感文本的判断需要预先建立敏感关键词库,关键词库由金融类网页以及钓鱼网页中具有典型业务特征的敏感文本组成,其作用是判断待测文本是否为敏感文本,判断步骤如下。

- 1) 使用网络爬虫工具获取待测页面。
- 2) 利用正则表达式提取页面 HTML 中 a、h、span 标签的字符数不超过长度 L 的文本信息。

3) 对文本进行预处理以去除空格等特殊字符, 得到不为空的文本条数 T_1 。

4) 将提取的文本与敏感文本库进行匹配, 如果提取的文本包含敏感文本, 则认为提取的文本是敏感文本; 统计得到敏感文本条数 $T_2 (T_2 \leq T_1)$ 。

2.2 logo 图像特征

金融类钓鱼网页总是会出现被仿冒金融机构的 logo 图像, 因此, 可以根据待测 logo 图像特征是否匹配金融类 logo 图像特征来对待测网页类型进行检测。logo 图像特征匹配方法的执行需要预先建立 logo 图像库, 图像库由金融类网页以及钓鱼网页的 logo 图像组成, 匹配步骤如下。

1) 使用浏览器工具获取待测页面, 自动调整页面尺寸为一固定大小。

2) 利用截图工具截取页面中的 logo 图像, 得到 logo 截图; 可以使用截图工具对网页的 logo 图像进行定点截图, 以网页左上角顶点为原点, 建立直角坐标系, 网页 logo 截图的位置示意如图 2 所示。

3) 使用图像特征提取算法分别对 logo 截图和预先建立的 logo 图像库中的图像 $i (i > 0)$ 进行特征点提取, 得到特征点个数分别记为 N_0 和 N_i 。

4) 计算两幅 logo 图像能够匹配的特征点个数, 记为 M_i 。

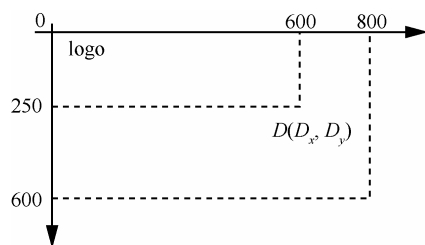


图2 网页 logo 截图的位置示意

截图的起点可以确定为原点, 只需确定截图终点 D 的坐标。使用网络爬虫随机爬取含有 logo 的网页若干例, 截取终点坐标分别为 $(D_x, 600)$ 和 $(800, D_y)$ 时的图片, 统计出能够完整包含 logo 的截图数量百分比 P_x 和 P_y , 得到 P_x 随 D_x 以及 P_y 随 D_y 的变化关系, 如图 3 所示。能够完整包含 logo 的截图尺寸越小, 就能尽量避免截取其他正常网页中含有的金融机构 logo, 同时图像特征匹配效率以及检测准确率也就越高, 根据图 3 选取终点 D 的坐标为 $(550, 280)$ 。

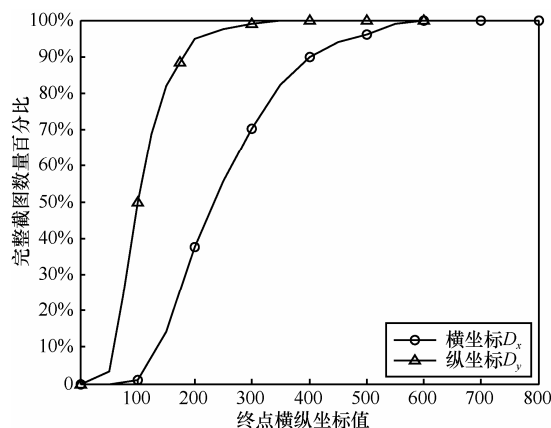


图3 完整截图数量百分比与终点横纵坐标值的关系

3 检测方法的构建

3.1 文本多模式匹配

网页 HTML 的 a、h、span 标签中文本长度一般都较短, 通过统计大量网页 3 种标签的文本长度, 确定一个适宜的文本长度值 L , 以提高敏感文本提取效率、降低钓鱼网页检测误判率。

在进行关键词模式串与文本串匹配时, 单模式匹配算法不仅能避免回溯, 而且还能利用特定的规则, 如坏字符规则和好后缀规则, 实现模式串的跳跃式匹配, 可提高匹配效率。但单模式匹配算法在处理多个模式串匹配时效率较低, 而多模式匹配算法扫描一遍文本串即能匹配所有模式串。经典的多模式匹配算法是基于有限状态自动机的 AC 算法^[9], 在处理大规模中文模式串匹配时, AC 算法及其改进算法需要的存储空间快速膨胀, 导致状态转移函数的计算量非常大, Cache 命中率降低, 算法的时间性能快速下降。

本文采用文献[10]提出的一种适合中文的多模式匹配算法 AC_SC, 该算法采用邻接链表存储有限状态自动机, 较好地解决了有限状态自动机存储空间快速膨胀问题。同时, 将状态为 0 的链表转化为散列链表, 以提高算法的时间性能, 算法描述如下。

1) 预处理阶段

①建立邻接表。建立顶点表 *VerticeTable*, 用于记录单链表的表头地址, 从状态 0 开始, 利用广度优先搜索建立的有限状态自动机能为跳转的下一状态建立一个结点, 并将其链接到 *VerticeTable*[0], 其中, 有限状态自动机的构造过程同 AC 算法一致。之后, 继续对下一状态进行处理, 直到所有

状态都完成处理。

②将状态0的单链表转换为散列链表。在状态0的链表中，为每一节点字符的首字节计算散列值，并将该节点链接到散列表中。

③计算状态转移函数 *goto*。若当前状态为0，根据标注字符首字节的散列值搜索散列表中对应的链表；否则，搜索 *VerticeTable[s]* 对应的单链表。

④计算失效函数 *failure*。对于状态 *s*，如果父状态为 *r*，标注字符为“*a*”，则 *failure(s)=goto(failure(r),a)*。

⑤计算输出函数 *output*。每处理完成一个关键词模式串，将该模式串加入到当前状态 *s* 的输出函数中。若 *failure(s) = s'*，则 *output(s) = output(s) ∪ output(s')*。

2) 匹配过程

①将当前状态 *s* 初始化为0，文本串指针指向文本串的首字符。

②若文本串指针不为空，则取出所指的字符“*a*”；否则结束匹配过程。

③计算 *s' = goto(s, a)*。若状态 *s* 为0，则根据字符的散列值搜索散列链表中对应的单链表；否则，搜索 *s* 对应的单链表。

④若 *s' = fail*，当前状态 *s = failure(s)*；若 *s* 为0，转步骤③。

⑤若 *s' ≠ fail*，则 *s ≠ s'*。如果 *output(s) =*

NULL，转步骤②；否则，直接输出 *output(s)*，表示关键词模式串匹配成功，转步骤②。

文本关键词采用邻接表存储方式。例如，关键词模式串集={建行，身份证号，转账，汇款，网银}，存储方式如图4所示。

通过匹配得到的敏感文本条数，可以计算文本中敏感文本的比例，即敏感文本特征值 *E* 为

$$E = \frac{T_2}{T_1} \quad (1)$$

其中，*T₁* 表示文本条数，*T₂* 表示敏感文本条数。

特征值 *E* 的大小反映待测网页在文本特征方面与金融类钓鱼网页的接近程度。如果 *E* 不小于阈值 *E_r*，则认为是金融类钓鱼网页，否则认为待测网页为可疑网页，需进一步进行网页 *logo* 相似度检测。

3.2 logo 图像特征提取

截取的网页 *logo* 图像与图像库中图像相对比，可能存在拉伸、背景改变以及 *logo* 颜色改变等问题，要求图像特征提取算法面对这些问题时具有较好的顽健性。典型的图像特征提取算法有 SIFT^[11]、SURF^[12] 以及 ORB^[13] 等，而 ORB 算法没有解决尺度不变性。SIFT 算法因具有很好的顽健性和抗干扰性，得到了大量应用。但因 SIFT 算子维度高（128 维），导致其时间复杂度较高。国内外研究者针对 SIFT 算子高维的缺点，不断对

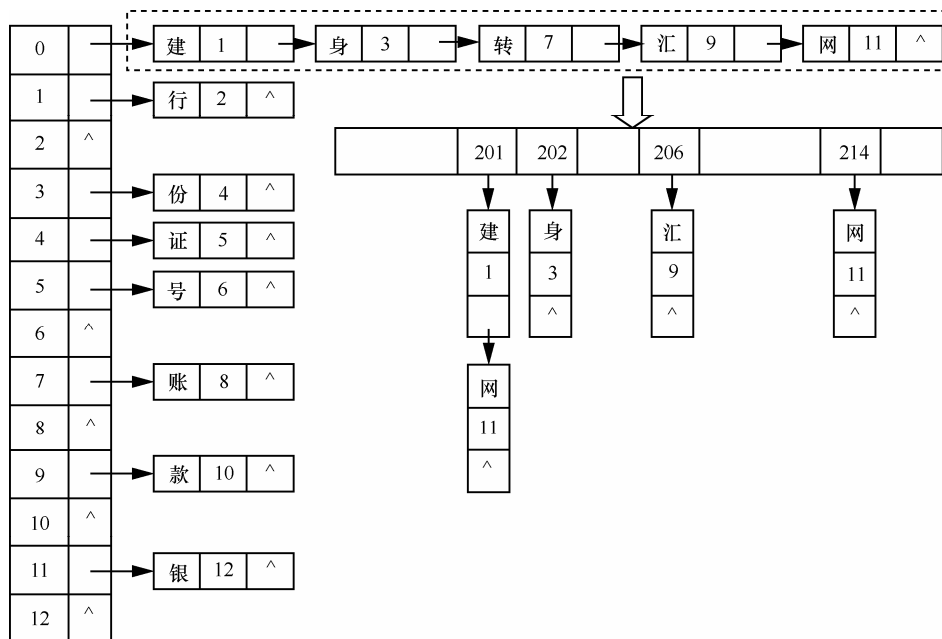


图4 关键词模式串的邻接表存储方式

其进行尝试和改进,以求在保持 SIFT 算子良好特性的前提下,尽量降低其维度。在这些研究中,具有代表性的是 PCA-SIFT 和 GLOH 算法,两者都采用了主成分分析(PCA, principal component analysis)技术。文献[14]对 SIFT、PCA-SIFT 以及 SURF 分别进行了时间、尺度、旋转、模糊、亮度等指标的对比,认为 PCA-SIFT 算法在各项指标中都能取得较好的表现,故本文利用 PCA-SIFT 算法作为 logo 图像特征匹配算法。

利用高斯卷积核是尺度变换的唯一线性核的特性,定义一幅 logo 图像的尺度空间 L 为

$$L(x, y, \sigma) = G(x, y, \sigma)I(x, y) \quad (2)$$

其中, $I(x, y)$ 表示 logo 图像在位置 (x, y) 处的像素值,二维高斯函数 G 为

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x-\frac{m}{2})^2 + (y-\frac{n}{2})^2}{2\sigma^2}} \quad (3)$$

其中, $m \times n$ 表示高斯模板的大小, m 和 n 的取值与 σ 相关; σ 是高斯分布的均方差,称为尺度空间因子,其值越小表示图像被平滑的越少。大尺度对应 logo 图像的概貌,小尺度对应图像的细节。利用尺度空间构建 logo 图像高斯金字塔,原始图像为金字塔的第一层,每次降采样所得到的图像为金字塔的第 i 层,金字塔的总层数 O 由 logo 图像原始尺寸 $P_0 \times Q_0$ 和塔顶图像的尺寸 $P_{O-1} \times Q_{O-1}$ 共同决定,总层数 O 为

$$O = \text{lb}\{\min(P_0, Q_0)\} - \text{lb}\{\min(P_{O-1}, Q_{O-1})\} \quad (4)$$

为了体现尺度的连续性,利用不同的参数 σ 对 logo 图像金字塔中每一层图像做高斯模糊,这样金字塔每层就有多张模糊图像,这些图像合称为一组或八度(octave)。金字塔中一组图像的初始图像是由前一组图像的倒数第 3 张图像隔点降采样得到。关键点是由空间的局部极值点组成,为了实现对图像关键点的初步探查,借助高斯差分空间,其表达式为

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma))I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \end{aligned} \quad (5)$$

实际计算过程中,高斯金字塔每组中相邻两层图像相减得到高斯差分金字塔。关键点由高斯差分空间的局部极值点组成,检测极值点需要每个像素点和它所有的相邻点比较,看其是否比它的图像域和尺度域的相邻点大或小。高斯金字塔

每组需要 L 层图像,高斯差分金字塔每组需 $L-1$ 层图像,就可以在每组中检测 $L-3$ 个尺度的极值点。构建尺度空间需要确定 3 个参数:尺度空间 σ 、组数索引 o 、组内层数索引 l ,这 3 个参数的关系如式(6)所示。

$$\sigma(o, l) = \sigma_0 2^{o+\frac{l}{L}}, o \in [0, \dots, O-1], l \in [0, \dots, L-1] \quad (6)$$

取式(5)中的 k 为组内总层数的倒数,即 $k = 2^{\frac{1}{L}}$,则式(5)可重新记为

$$D(x, y, \sigma) = L(x, y, \sigma(o, l+1)) - L(x, y, \sigma(o, l)) \quad (7)$$

为了提高关键点的稳定性,在获得极值点后对其进行筛选,去掉具有强烈边缘效应和低对比度的候选点,以增强匹配的稳定性和抗噪能力。对尺度空间高斯差分函数进行二阶泰勒展开,有

$$D(X) = D + \frac{\partial D^T}{\partial X} X + \frac{1}{2} X^T \frac{\partial^2 D}{\partial X^2} X \quad (8)$$

其中, $X = (x, y, \sigma)^T$ 是偏移量,令式中一阶导数为 0,可以得到极值点相对于插值中心的偏移量,为

$$\hat{X} = -\frac{\partial^2 D^{-1}}{\partial X^2} \frac{\partial D}{\partial X} \quad (9)$$

利用在偏移向量 $\hat{X}$ 处约束 $|D(\hat{X})|$ 小于某一阈值,可以去除对噪声敏感的特征点,也可以滤掉对比度低的特征点。将式(9)代入式(8)中,有

$$D(\hat{X}) = D + \frac{1}{2} \frac{\partial D^T}{\partial X} \hat{X} \quad (10)$$

对边缘敏感的特征点通常在高斯函数边缘某一方向的曲率值相对较大,而在另一个方向的曲率值会相对较小,运用 2×2 的 Hessian 矩阵 H 作为稳定性判别标准,剔除不稳定的关键点以增强匹配稳定性,提高抗噪声能力。为了使描述符具有旋转不变性,以关键点为中心的邻域窗口内采样,并用直方图统计邻域像素的梯度方向,直方图的峰值则代表了该关键点处邻域梯度的主方向,即作为该关键点的主方向。PCA-SIFT 算法保留了 SIFT 中尺度空间极值检测、特征点亚像素定位、主方向分配等主要步骤^[15],主要对特征描述子生成方式进行改进,步骤如下。

1) 以关键点为中心选择 41×41 像素的像斑。

2) 计算像斑内每个像素水平和垂直方向的梯度,形成一个大小为 $2 \times 39 \times 39 = 3\,042$ 维的输

入向量。

3) 所有关键点的输入向量组成一个 $b \times 3\ 042$ 维的矩阵 A , 其中, b 为关键点的个数。

4) 计算矩阵 A 的协方差: $A = A - \text{mean}(A)$, $\text{cov}(A) = A^T A$; 计算协方差矩阵的特征值 λ 与特征向量 β 。

5) 将特征值从大到小排列, 选择前 j 个特征值对应的特征向量组成一个 $j \times 3\ 042$ 维的投影矩阵 B , j 既可以根据经验设置的固定值, 也可以动态选择。

6) 将原矩阵 A 与投影矩阵 B 相乘, 得到新的 j 维 PCA-SIFT 特征点描述子。

待测 logo 图像和预先建立的 logo 图像库中图像 i 经过 PCA-SIFT 算法处理后, 可以分别生成两幅 logo 图像的特征点描述子 N_0 和 N_i , 分别是 k_1 和 k_2 维。取待测 logo 图像中的某个特征点, 并找出其与图像 i 中欧式距离最近的前 2 个特征点。在这 2 个特征点中, 如果最近距离与次近距离的比值小于某一阈值, 则认为这 2 个匹配点匹配。计算出 logo 图像匹配的特征点总个数 M_i , 从而可以得到两幅 logo 图像的相似度 S 为

$$S = \frac{M_i}{N_i} \quad (11)$$

相似度 S 的大小可反映待测网页与金融类网页的相似程度。如果 S 不小于阈值 S_T , 则认为是金融类钓鱼网页, 否则认为待测网页为正常网页。为提高图像特征匹配效率, 可用 PCA-SIFT 算法在钓鱼网页检测之前提取图像库中的所有图像的特征点。

3.3 检测方法的实现

在利用本文检测方法之前, 待测 URL 需要先分别经过域名白名单、域名黑名单以及网络访问量排名的过滤, 这样可以过滤掉正常的金融网页和大部分常用网页, 提高检测效率。域名白名单由正常的金融网页域名以及常用的网页域名组成, 域名黑名单由已经确认的钓鱼网页域名组成, 域名白名单、黑名单具有“新陈代谢”机制, 使用 Alexa 过滤 URL 网络访问量排名在 100 万以内的 URL。

根据上述方法, 采用 Python 语言实现了金融类钓鱼网页检测系统, 该系统具备线程安全, 可

批量接收含有可疑 URL 的文本。使用者可以在检测系统中批量添加敏感文本和金融类网页 logo 图像。在敏感文本特征值和图像相似度的计算过程中, 为提高钓鱼网页的检测效率, 当 $E \geq E_T$ 或者 $S \geq S_T$ 时, 则停止后续计算并得出检测结果。基于页面敏感特征的金融类钓鱼网页检测流程如图 5 所示。

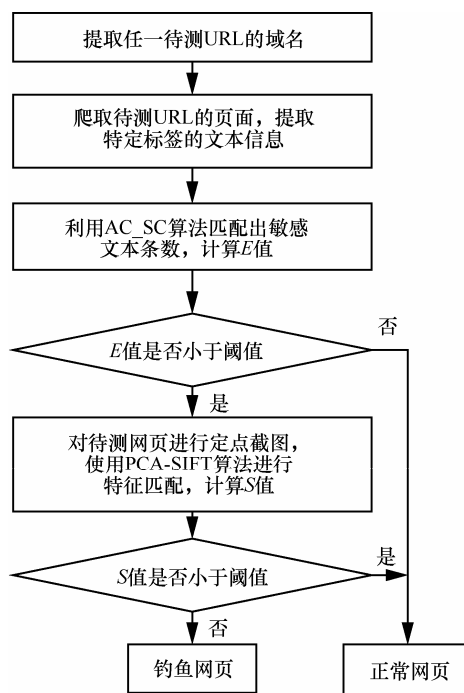


图5 钓鱼网页检测流程

4 实验结果与分析

为验证本文所构建金融类钓鱼网页检测系统的识别效果, 需要收集钓鱼网页作为测试用例。收集 2016 年 9 月 1 日至 2016 年 10 月 31 日 PhishTank 中被举报的钓鱼网页 URL, 随机选取并进行验证, 得到经确认的金融类钓鱼网页 URL 有 135 例。使用谷歌搜索在线服务的网站, 随机选取搜索的结果并进行验证, 共收集 150 例合法的网站 URL。将 285 个测试用例随机混合在一起, 使用本文的检测方法进行多线程检测。为了测试本文方法的有效性, 选取多个与本文相似的钓鱼网页检测的方法进行比较。文献[16]和文献[17]中钓鱼网页检测方法与本文都是基于页面敏感特征, 文献[18]是基于登录页面和 logo 图像实现钓鱼网页的检测, 对比结果如表 1 所示。

表1 钓鱼网页检测结果比较

方法	样本数/个	召回率	误报率
文献[16]	100	94.9%	N/A
文献[17]	100	90.0%	2.0%
文献[18]	1 000	93.6%	1.4%
本文方法	285	97.8%	3.3%

通过实验结果可以发现, 本文方法召回率最高, 可达到 97.8%, 文献[18]的误报率最低。对于误判的合法网页, 经过分析, 发现它们的页面中 a 标签、h 标签以及 span 标签含有较多的金融类敏感文本, 造成误判。而对于漏报的钓鱼网页, 通过查看网页源代码发现, 页面中含有极少的文本且没有 logo 图像, 造成漏报。考虑到本文方法针对的是金融类钓鱼网页, 文献[16~18]中的方法是针对所有的钓鱼网页, 如果重点检测金融类钓鱼网页, 其他 3 个方法达不到现有的召回率和误报率指标, 说明本文方法对金融类钓鱼网页具有很好的识别效果。

5 结束语

本文通过对大量金融类钓鱼网页样本的分析, 在传统的基于黑白名单和页面相似度检测方法的基础上, 总结出表征金融类页面的敏感特征以及描述方法, 提出基于敏感特征的金融类钓鱼网页检测方法, 并进行系统实现。检测系统利用多模式匹配算法 AC_SC 匹配出网页中的敏感文本, 利用 PCA-SIFT 算法提取 logo 图像的特征, 结合敏感文本特征值和 logo 图像相似度, 实现对金融类钓鱼网页的检测和判定。实验结果表明, 本文方案敏感特征的提取对于绝大部分钓鱼网页的判定是有效的, 可取得较高的召回率和较低的误报率。

参考文献:

- [1] 中国反钓鱼网站联盟. 2016 年 1-10 月钓鱼网站处理简报[R]. 北京: APAC, 2016.
- [2] 亿盾互联. 亿盾互联 2016 年 10 月钓鱼网站检测月报[EB/OL]. http://www.yeeton.com.cn/report_month.asp.htm.
- [3] 沙泓洲, 刘庆云, 柳厅文, 等. 恶意网页识别研究综述[J]. 计算机学报, 2016, 39(3): 529-542.
- [4] SHA H Z, LIU Q Y, LIU T W, et al. Survey on malicious webpage detection research[J]. Chinese Journal of Computers, 2016, 39(3): 529-542.
- [5] PRAKASH P, KUMAR M, KOMPPELLA R R, et al. PhishNet: predictive blacklisting to detect phishing attacks[C]//IEEE Infocom. 2010: 1-5.
- [6] ZHANG Y, HONG J I, CRANOR L F. CANTINA: a content-based approach to detecting phishing Web sites[C]//The International Conference on World Wide Web. 2007: 639-648.
- [7] XIANG G, HONG J, ROSE C P, et al. CANTINA+: a feature-rich machine learning framework for detecting phishing Web sites[J]. ACM Transactions on Information and System Security, 2011, 14(2): 613-613.
- [8] ZHANG W F, ZHOU Y M, XU L, et al. A method of detecting phishing Web pages based on hungarian matching algorithm[J]. Chinese Journal of Computers, 2010, 33(10): 1963-1975.
- [9] SHENG S, WARDMAN B, WARNER G, et al. An empirical analysis of phishing blacklists[C]//The 6th Conference on Email and Anti-Spam. 2009: 59-78.
- [10] AHO A V, CORASICK M J. Efficient string matching: an aid to bibliographic search[J]. Communications of the ACM, 1975, 18(6): 333-340.
- [11] 候整风, 杨波, 朱晓玲. 一种适合中文的多模式匹配算法[J]. 计算机科学, 2013, 40(11): 117-121.
- [12] HOU Z F, YANG B, ZHU X L. Multiple pattern algorithm for chineses[J]. Computer Science, 2013, 40(11): 117-121.
- [13] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.
- [14] BAY H, TUYTELAARS T, GOOL L V. SURF: speeded up robust features[J]. Computer Vision & Image Understanding, 2006, 110(3): 404-417.
- [15] RUBLEE E, RABAUDE, KONOLIGE K, et al. ORB: an efficient alternative to SIFT or SURF[C]//IEEE International Conference on Computer Vision. IEEE, 2012: 2564-2571.
- [16] LUO J, GWUN O. A comparison of SIFT, PCA-SIFT and SURF[J]. International Journal of Image Processing, 2009, 3(4): 143-152.
- [17] KE Y, SUKTHANKAR R. PCA-SIFT: a more distinctive representation for local image descriptors[C]//The 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2004: 506-513.
- [18] 王燕, 王兴芬, 任俊玲. 面向钓鱼网站敏感特征项选取的 IIGAIN 算法[J]. 计算机应用于软件, 2016, 33(4): 297-301.
- [19] WANG Y, WANG X F, REN J L. IIGAIN algorithm oriented to phishing websites sensitive feature items selection[J]. Computer Applications and Software, 2016, 33(4): 297-301.
- [20] 宋明秋, 曹晓芸. 基于敏感特征的网络钓鱼网站检测方法[J]. 大连理工大学学报, 2013, 53(6): 903-907.
- [21] SONG M Q, CAO X Y. Phishing detection method based on sensitive characteristics of phishing webpage[J]. Journal of Dalian University of Technology, 2013, 53(6): 903-907.
- [22] YANG M X, PENG Y B, PENG X G. A novel anti-phishing scheme based on login page detection and logo identification[J]. Journal of Computational Information Systems, 2004, 10(23): 10179-10187.

作者简介:



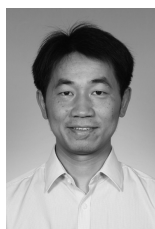
胡向东 (1971-), 男, 四川广安人, 博士, 重庆邮电大学教授, 主要研究方向为网络化测控及其信息安全、物联网与智慧空间安全、复杂系统建模仿真与优化。



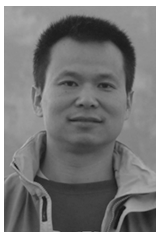
林家富 (1989-), 男, 四川成都人, 重庆邮电大学硕士生, 主要研究方向为物联网安全。



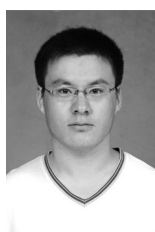
刘可 (1992-), 男, 重庆人, 重庆邮电大学硕士生, 主要研究方向为物联网安全。



付俊 (1979-), 男, 湖北松滋人, 中国移动通信有限公司研究院研究员, 主要研究方向为网络与信息安全方案设计、安全标准制定以及黑客攻防对抗技术。



张峰 (1977-), 男, 湖北孝感人, 博士, 中国移动通信有限公司研究院高级工程师, 主要研究方向为网络与信息安全技术应用。



郭智慧 (1986-), 男, 河北张家口人, 中国移动通信有限公司研究院研究员, 主要研究方向为网络欺诈治理。