

# 基于嵌套 EMD 的钓鱼网页检测算法

曹玖新 毛 波 罗军舟 刘 波

(东南大学计算机科学与工程学院 南京 210096)

**摘 要** 网络钓鱼(Web phishing)以相似网站欺诈用户、骗取个人机密信息,已成为电子金融活动的重大威胁.对此,文中提出了一个钓鱼网页检测架构.在具体检测机制方面,提出了一个基于嵌套 EMD(Nested Earth Mover's Distance)的网页相似度判定算法,对 Web 图像进行分割,抽取子图特征并构建网页的 ARG(Attributed Relational Graph),在计算不同 ARG 属性距离的基础上,采用嵌套 EMD 方法计算网页的相似度,实现了对钓鱼网站的检测.实验结果表明,与国际现有研究成果相比,该算法具有较高的精度和较强的适应性.

**关键词** 钓鱼网页检测;网页分割;特征关系图;嵌套 EMD;相似度

**中图法分类号** TP393 **DOI 号**: 10.3724/SP.J.1016.2009.00922

## A Phishing Web Pages Detection Algorithm Based on Nested Structure of Earth Mover's Distance (Nested-EMD)

CAO Jiu-Xin MAO Bo LUO Jun-Zhou LIU Bo

(School of Computer Science and Engineering, Southeast University, Nanjing 210096)

**Abstract** Web Phishing has become a big threat to online applications such as financial services, it steals user identities and credentials by imitating the sites of service providers such as banks. This paper proposes a novel architecture of Phishing Web detection which gives the function modules and processing workflow, and a visual based Web page similarity detecting algorithm. Based on the image of the suspicious Web page, the algorithm first divides Web page into sub-block images from which features and relations are abstracted and the ARG (Attributed Relational Graph) of the Web page is formed. Then based on the ARG of two Web pages, we get the Nested-EMD (Earth Mover's Distance) of the two pages as their similarity, and then the decision can be concluded by comparing the similarity degree between two Web pages. The algorithm is implemented and compared with the latest international researches, and it is shown that the algorithm is better in accuracy and robustness according to the experiment.

**Keywords** phishing detection; Web segmentation; attributed relational graph; nested EMD; similarity degree

收稿日期:2007-10-22;最终修改稿收到日期:2009-01-21.本课题得到国家自然科学基金(90604004,60773103)、高校博士点专项科研基金(200802860031)、江苏省自然科学基金(BK2007708,BK2008030)、江苏省“网络与信息安全”重点实验室(BM2003201)和“计算机网络和信息集成”教育部重点实验室(93K-9)资助.曹玖新,男,1967年生,博士,副教授,主要研究方向为计算机网络安全、服务计算和数字版权管理. E-mail: jx.cao@seu.edu.cn.毛 波,男,1982年生,博士研究生,主要研究方向为计算机网络安全、数字城市.罗军舟,男,1960年生,博士,教授,博士生导师,主要研究领域为下一代网络体系结构、协议工程、网络安全和管理、网格计算.刘 波,女,1975年生,博士,讲师,主要研究方向为普适计算和服务管理.

## 1 引言

网络钓鱼是指那些利用与原网页极其相似的假冒网页骗取用户个人信息(如银行帐号、密码等)的行为。随着电子商务等网络应用的快速发展,网络钓鱼的危害逐年增加<sup>①</sup>,这引起了产业界和学术界广泛的关注,并提出了一系列的防范措施。

现有的反网络钓鱼技术可以分为 3 大类:基于服务器的防范、基于浏览器的防范和独立的第三方检测。基于服务器的防范指服务器通过认证来防范网络钓鱼,例如电子证书、动态安全皮肤<sup>[1]</sup>等。基于浏览器的防范措施通过嵌入浏览器的插件来提示用户。独立的第三方防范措施主要目的是发现并共享钓鱼网站相关信息,包括电子邮件检测<sup>[2]</sup>、网络行为检测<sup>[3]</sup>、个人信息保护<sup>[4]</sup>、网页异常检测<sup>[5]</sup>、实时黑名单以及网页相似性检测等<sup>[6-7]</sup>。由于钓鱼网站能绕过服务器,基于服务器的措施无法有效地防范网络钓鱼。基于浏览器的措施需要第三方提供的钓鱼网站黑名单,因此,钓鱼网站的检测是防范网络钓鱼的基础。但由于网络钓鱼的复杂性,仅使用单一的检测防范措施难以达到预期效果,对此本文提出了一套完整的钓鱼网页检测体系架构,并深入研究了钓鱼网页检测的核心算法——网页相似性检测算法。

本文所提出的钓鱼网站检测体系包括垃圾邮件检测、网络钓鱼分析节点以及网络钓鱼控制中心三个部分。网页相似性检测算法则包括网页图像的分割、特征抽取、位置关系向量矩阵形成、子块关系(ARG)生成以及嵌套 EMD 距离的计算等步骤,实验验证该算法可以有效地检测出两个网页的相似性。

本文第 2 节给出了国内外的相关研究现状;第 3 节给出钓鱼网页检测体系构架;第 4 节介绍网页相似性检测算法;实验结果在第 5 节给出;最后对我们的工作进行总结。

## 2 相关工作

目前,钓鱼网页检测研究集中在网页异常检测、网络行为检测以及基于视觉的钓鱼网页检测等几个方面。

Pan Ying 等<sup>[5]</sup>提出了一种基于网页异常的检测,该方法基于网页的 DOM 结构,使用 SVM(Support Vector Machine)检测钓鱼网页,但该方法无法处理网页中的图片,从而大大降低了算法的准确性。

Madhusudhanan 等<sup>[3]</sup>则通过模拟用户的行为

检测钓鱼网站,但该方法无法防范桥接攻击和网站的机器人检测手段。

基于视觉的检测分为基于 HTML 文本的匹配和基于图像的匹配。由于 HTML 语言的灵活性以及网页元素的动态性和丰富性,仿冒者可以轻易地做出看上去一样但 HTML 结构完全不同的网页,对此基于 HTML 的匹配将完全失效。而基于图像的网页相似检测方法根据人的视觉原理,对网页的视觉相似度进行判定,因而是一种高效和通用的检测方法<sup>[6]</sup>。Cordero 等<sup>②</sup>提出了一种使用 SVM 的网页图像检测算法,但该方法只能用于某个网站的检测,同时数学特性十分复杂。Fu 等提出了一种基于像素及其位置的 EMD 距离的匹配算法,从其实验结果可以看出效果要明显好于基于 HTML 内容的检测,但该算法只考虑了网页图像中的颜色及其分布特点,没有考虑网页中不同部分之间的位置关系,根据格斯塔视觉原理<sup>[8]</sup>,相对位置在人的视觉中占主要地位,特别是多个形体间的相对位置关系,相对位置关系的变化必然导致视觉上的区别,而该算法由于没有考虑相对位置因素可能导致相似检测的失效。

针对相关工作的不足,本文提出了一种高效钓鱼网页检测算法——基于图像分割和嵌套 EMD 的钓鱼网页检测算法,该方法通过对网页图像进行分割、子图特征提取、嵌套 EMD 距离计算等步骤构建网页的 ARG(Attribute Relation Graph),从而对其进行匹配计算并获得可疑网页与受保护网页直接的视觉相似度,最终完成钓鱼网页的检测判定。

EMD(Earth Mover's Distance)是一种用于判断两个特征集之间距离的数学方法,该方法源自著名的运输问题。而嵌套 EMD 则是 Kim 等提出的一种图的匹配算法<sup>[9]</sup>,该算法可以更有效地处理多维特征向量并具有很高的抗噪性能。

## 3 钓鱼网页检测体系

随着网络钓鱼的国际化、专业化,要应对该威胁,必须找到一个能联合各方力量(包括研究机构、政府、银行、服务提供商、用户等)并基于现有安全基础设施的易部署、可管理的网络钓鱼防范体系。

基于以上思路,我们提出了钓鱼网站检测体系架构,如图 1 所示。

① <http://www.apwg.org>, APWG Report January 2007

② <http://www.cs.berkeley.edu/~asimma/294-fall06/projects/reports/corder.pdf>



$y_1), P_2(x_2, y_2)$ ——图片的边界点,迭代进行以下 5 个步骤,直到所有的子图都无法再分割。

#### ① 收缩。

检测出非 0 像素的边界,即求出包含所有非 0 像素的最小矩形区域  $P_{\min}(x_{\min}, y_{\min}), P_{\max}(x_{\max}, y_{\max})$ ;

#### ② 判断该区域是否需要继续分割。

如果该区符合太窄或太短则不进行分割,并将  $P_{\max}$  和  $P_{\min}$  记为 BW 的一个块,否则继续分割;

#### ③ 检测分割带。

分割带是指可能将该区域分为两部分的区域,根据网页的特点有水平和垂直两种。分割带既可以由 0 构成(代表背景),也可以由 1 构成(代表边界)。首先在  $P_{\max}$  和  $P_{\min}$  之间,分别在水平和垂直方向上进行检测,如果某一行或列的构成趋于一致(绝大部分像素都为 0 或 1),则该行或列为一个分割带;然后合并分割带,分别在水平和垂直方向将相邻

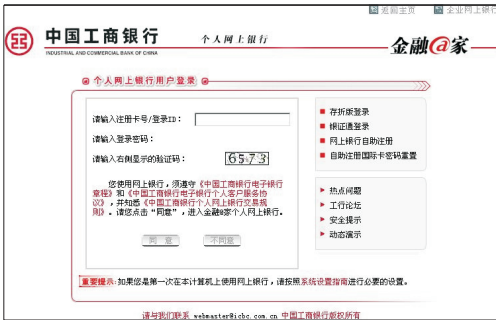
的分割带合并成一个具有一定宽度的分割带。如果未检测出分割带则说明该部分不可分割,并将  $P_{\max}$  和  $P_{\min}$  记为 BW 的一个部分。

#### ④ 选择分割带。

如果检测出了分割带则可根据某种策略选择其中之一作为依据分割 BW。选择的策略一般为:优先选择边界分割带;优先选择宽背景分割带;如果当前图像为长条状则优先选择水平分割带,否则选择垂直分割带;

⑤ 根据选择的分割带将图片分为两部分,并对这两部分分别重复 a~e 的检测过程。

经过以上算法,得到的关键区域均为矩形,这也符合网页中基本元素都为矩形的特点。如果需要检测边界内部,则可在收缩时将边界除去,网页分割效果如图 2 所示。



(a) 分割前



(b) 分割后

图 2 网页分割

## 4.2 特征图的生成

网页的特征关系图(ARG)包括组成该网页每个块的特征(本文采用彩色、灰度直方图以及长宽表示)和块之间的相对位置关系(使用一个 9 维向量描述)。首先对网页中的每个块提取其特征(彩色、灰度直方图和长宽),再根据块的位置分布,计算出块与块之间的相对位置关系,具体如以下小节。

### 4.2.1 关键区域特征提取

块的特征由一个特征向量表示  $V = \{S, H, G\}$ , 其中  $S = \{w, h\}$  为区域边界形状(包括长和宽),  $H$  为彩色直方图,  $G$  为灰度直方图。其中边界形状用于描述区域内形状相关的特性,彩色和灰度直方图用于描述颜色信息,具体计算如下:

(1) 边界形状  $S$ . 由分割的结果直接给出  $S = \{w, h\}$ , 本文采用长和宽表示。

(2) 彩色直方图  $H$ . 将原图从 RGB 空间转换至 HSV 空间,并将 HSV 空间非均匀量化为 32 种颜色,具体量化方法参见文献[10],彩色直方图特征向量  $H$  的定义由式(1)给出:

$$H = \langle h[c_1], h[c_2], \dots, h[c_k], \dots, h[c_n] \rangle,$$

$$\sum_{k=1}^n h[c_k] = 1, 0 \leq h[c_k] \leq 1 \quad (1)$$

其中  $h[c_k]$  表示第  $k$  种颜色像素的频数,即

$$h[c_k] = \frac{\sum_{i=1}^h \sum_{j=1}^w \begin{cases} 1, & Q(i, j) = c_k \\ 0, & \text{其它} \end{cases}}{\text{width} \times \text{height}} \quad (2)$$

(3) 灰度直方图  $G$ . 将原图由 RGB 空间转换到灰度空间,并将灰度空间量化为 32 个灰度等级,量化公式为  $v = (v_0 \times 32) / (v_{\max} - v_{\min} + 1)$ , 其中  $v_0$  是原灰度等级,  $v_{\max}$  是该子图中最大灰度,  $v_{\min}$  为最小灰度,  $v$  为量化后灰度。量化后统计各灰度出现的频数,得到灰度直方图特征向量  $G$ ,

$$G = \langle g[v_1], g[v_2], \dots, g[v_k], \dots, g[v_n] \rangle,$$

$$\sum_{k=1}^n g[v_k] = 1, 0 \leq g[v_k] \leq 1 \quad (3)$$

其中  $g[v_k]$  表示第  $k$  种像素的灰度频数,即

$$g[v_k] = \frac{\sum_{i=1}^h \sum_{j=1}^w \begin{cases} 1, & I(i, j) = v_k \\ 0, & \text{其它} \end{cases}}{\text{width} \times \text{height}} \quad (4)$$

至此已经取得了子图的特征值  $v$ , 其中  $v = \{size, H, G\}$ . 将所有子图的特征值组成特征向量

$V$ , 作为图  $G$  的结点特征向量.

#### 4.2.2 关系矩阵的生成

根据格斯塔理论<sup>[8]</sup>, 相对位置在视觉识别中占主要地位, 因此把关键区域(块)之间的相对位置作为关系矩阵生成的主要依据. 由分割算法可知本文中关键区域均为矩形, 以该矩形为中心, 将 2 维平面分为 9 个部分, 再求出另一区域在这 9 部分中的分布, 则可求出它们的相对位置关系. 具体计算过程如下图, 假设要求关键区域(块)  $KA_i$  与  $KA_j$  之间的相对位置关系  $r_{ij}$ , 先根据  $KA_i$  将平面分割成 9 部分, 再求出  $KA_j$  在这 9 个区域中的分布, 即  $r_{ij}[k] = (KA_{ik} \wedge KA_j)$ , 其中  $KA_{ik}$  表示  $KA_i$  的第  $k$  个区域. 为了简化计算, 设  $KA_{ik} \wedge KA_j = 1$  如果有共同区域, 否则为 0. 这样图 3 中所示的  $r_{ij} = \{0, 0, 0, 0, 1, 1, 0, 0, 0\}$ , 由于只有  $KA_{i5}, KA_{i6}$  与  $KA_j$  有共同区域, 所以  $r_{ij}$  的第 5 和第 6 个分量为 1, 其余全为 0. 特别的对于任何  $i, r_{ii} = \{0, 0, 0, 0, 0, 0, 0, 0, 1\}$ .

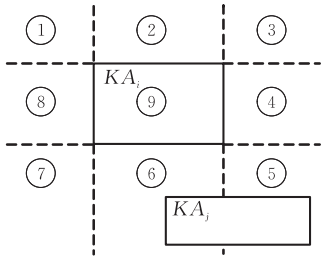


图 3 关键区域(块)的相对位置关系

求出所有子图之间的关系向量并生成关系矩阵  $R = [r_{ij}]$ , 其中  $R$  为  $n \times n$  矩阵,  $n$  为子图个数,  $R$  的每个元素  $r_{ij}$  为一个 9 维向量, 该向量由 0 或 1 组成. 将  $R$  作为图  $G$  的关系矩阵, 这样就生成了网页图片的特征图  $G = \langle V, R \rangle$ .

#### 4.3 使用 EMD 嵌套结构计算相似度

每个网页用图表示为  $G = \langle V, R \rangle$ , 计算该图与被保护网页图  $G' = \langle V', R' \rangle$  之间的 Nest-EMD 距离分为两步, 首先算出内部 EMD 距离, 依据内部 EMD 形成外部距离矩阵, 从而求出这两个图之间的距离.

若要比较的两个图分别为  $G = \langle V, R \rangle, G' = \langle V', R' \rangle$ , 其中

$$V = \{v_i | 1 \leq i \leq n\},$$

$$R = \{r_{ij} | i! = j, 1 \leq i \leq n, 1 \leq j \leq n\},$$

$$V' = \{v'_i | 1 \leq i \leq m\},$$

$$R' = \{r'_{ij} | i! = j, 1 \leq i \leq m, 1 \leq j \leq m\}.$$

$v, v'$  为特征值, 定义参见 4.2.1 节,  $r_{ij}, r'_{ij}$  为关系向量, 定义参见 4.2.2 节.

#### 4.3.1 结点距离函数

为了计算特征值结点  $v$  和  $v'$  之间的距离, 可以对  $v$  和  $v'$  中各个分量之间的距离分别进行计算并将其加权和作为特征值结点距离. 同时为计算方便对距离进行了归一化处理.

(1)  $size$  间的距离. 由于  $size$  由宽和高两个分量组成, 设  $size_1 = \{w_1, h_1\}, size_2 = \{w_2, h_2\}, w_{\max} = \max(w_1, w_2), h_{\max} = \max(h_1, h_2), w_{\min} = \min(w_1, w_2), h_{\min} = \min(h_1, h_2)$ , 则  $size$  间距离  $d_{size} = 1 - (w_{\min} \times h_{\min}) / (w_{\max} \times h_{\max})$ , 该距离一方面计算简便, 同时能更加有效地反映面积的区别.

(2) 彩色直方图距离. 根据文献[11], 任意两个  $N$  维直方图  $H_p$  和  $H_q$  的相似度  $S_H(p, q)$  为

$$S_H(p, q) = \sum_{i=1}^N \min(H_p(i), H_q(i)),$$

$$\sum_{i=1}^N H_p(i) = \sum_{i=1}^N H_q(i) = 1, N=32 \quad (5)$$

其中  $H_p(i)$  表示彩色直方图  $p$  中彩色  $i$  的概率.

(3) 灰度直方图距离. 与彩色直方图类似, 任意两个  $N$  维灰度直方图  $G_p$  和  $G_q$  的相似度  $S_G(p, q)$  为

$$S_G(p, q) = \sum_{i=1}^N \min(G_p(i), G_q(i)),$$

$$\sum_{i=1}^N G_p(i) = \sum_{i=1}^N G_q(i) = 1, N=32 \quad (6)$$

且  $d_{size}, S_H$  和  $S_G$  都属于  $[0, 1]$ , 即为归一化数据.

将这 3 个数据加权相加后即得结点距离  $d(v_i, v'_i) = a \times d_{size} + b \times S_H + c \times S_G$ , 其中  $a + b + c = 1$ .

#### 4.3.2 关系距离函数

该函数反映  $r_{ij}$  和  $r'_{i'j'}$  之间的距离. 由于  $r$  为 9 维向量, 这里使用 EMD 距离计算  $r$  和  $r'$  之间的距离. 根据图 3, 任意两个区域之间都存在一个唯一的曼哈顿距离, 例如区域 5 和 1 之间的距离为 4, 以该距离作为基础可构成一个  $9 \times 9$  的距离矩阵. 基于该矩阵, 计算出任意两个关系向量  $r$  和  $r'$  之间的 EMD, 以此 EMD 距离作为  $r$  和  $r'$  之间的关系距离的一部分  $d_{EMD}(r, r')$ . 两个向量中非 0 元素个数之差作为关系距离的另一部分  $d_N(r, r') = |N(r) - N(r')|$ , 其中  $N(r)$  表示  $r$  中非零元素的个数. 则  $d(r, r') = (d_{EMD}(r, r') + d_N(r, r')) / d_{\max}$ , 其中  $d_{\max}$  为可能的最大关系距离.

#### 4.3.3 内部 EMD 距离的计算

内部 EMD 距离表示  $G$  和  $G'$  中给定的两个结点之间的 EMD 距离, 该距离可作为外部 EMD 矩阵中这两点之间的距离. 给定  $i$  和  $i'$  分别为  $G$  和  $G'$  中



的第  $i$  和第  $i'$  个结点,通过求内部距离矩阵的方法可以求出结点  $i$  和  $i'$  的内部 EMD. 内部距离矩阵  $\mathbf{D}_{\text{inner}}$  是一个  $n \times m$  矩阵,其中的元素  $d_{\text{inner}}(j, j') = (1-a) \times d(v_j, v_{j'}) + a \times d(r_{ij}, r_{i'j'})$ , 其中  $j$  属于  $[1, n]$ ,  $j'$  属于  $[1, m]$ ,  $a$  在  $[0, 1]$  之间,结点和关系的距离函数  $(v_j, v_{j'})$  由 4.3.1 节给出,  $d(r_{ij}, r_{i'j'})$  则由 4.3.2 节给出. 求出  $\mathbf{D}_{\text{inner}}$  后,以  $\mathbf{D}_{\text{inner}}$  为距离矩阵,以  $S = \{(w_j) | 1 \leq j \leq n\}$ ,  $S' = \{(w_{j'}) | 1 \leq j' \leq m\}$  为特征向量,使用 EMD 算法算出外部距离记作  $d_{\text{out}}(i, i')$ , 其中  $w_j = w_{j'} = 1/\max(n, m)$ , 可以更加有效地进行子图匹配.

求出所有的  $d_{\text{out}}$  后得到  $\mathbf{D}_{\text{out}} = [d_{\text{out}}(i, i')]$ . 并根据  $\mathbf{D}_{\text{out}}$  和  $S, S'$  (定义同上) 求出转移矩阵  $\mathbf{F}$  和外部 EMD 距离  $P$  作为  $G$  和  $G'$  之间的最终相似度,该距离越小说明  $G$  和  $G'$  越相似.

#### 4.4 算法复杂度分析

整个算法分为图像分割、特征提取以及 NEMD 距离计算 3 个步骤. 最坏情况下其复杂度为  $O(n \times w_0 \times h_0)$ , 其中  $n$  为分割后所得的子图个数,  $w_0, h_0$  为网页图像的宽和高(单位为像素), 由于一般情况下  $n < 20$  所以该算法的复杂度是可以接受的. 具体分析过程如下.

##### 4.4.1 图像分割复杂度

设原始网页图像的长和高分别为  $w_0, h_0$ , 子图  $i$  的长和高为  $w_i, h_i$ . 图片预处理算法 Canny 的时间复杂度为  $O(w_0 \times h_0)$ . 迭代分割中对子图像  $i$  进行收缩处理时只需对子图像所有像素进行一次扫描即可求出, 因此复杂度为  $O(w_i \times h_i)$ ; 判断是否能继续分割只需要进行一次比较操作, 复杂度为  $O(1)$ ; 分割带检测由于要扫描整个子图像, 并对扫描结果进行合并, 所以其时间复杂度也为  $O(w_i \times h_i)$ . 同类分割带(都为水平或垂直)的比较可以在检测中进行, 因此只需对水平和垂直分割带比较, 所以该操作的复杂度为  $O(1)$ ; 子图像的分割只需按照分割带与其位置坐标生成新的子图, 复杂度为  $O(1)$ . 由于子图像之间无重复, 有  $\sum w_i \times h_i < w_0 \times h_0$  (其中子图  $i$  属于待分割集合), 因此每迭代分割一次的时间复杂度小于  $O(w_0 \times h_0)$ , 设最后生成  $n$  个不可分割子图, 则最多需要  $n-1$  次分割(如果某次分割没有新的子图产生则根据算法可知分割结束), 所以整个分割算法的时间复杂度小于  $O(n \times w_0 \times h_0)$ , 其中  $n$  为分块个数.

##### 4.4.2 特征关系图生成复杂度

由于特征由子图像的彩色直方图、灰度直方图

以及长宽构成, 只需对每个子图进行一次扫描, 因此特征提取的复杂度为  $O(w_0 \times h_0)$ .

关系矩阵为  $n \times n$  的矩阵, 每计算一个元素的复杂度为定值所以生成关系矩阵的复杂度为  $O(n \times n)$ .

##### 4.4.3 NEMD 复杂度

设待匹配的两个网页中分别有  $n, m$  个子图, 其 NEMD 包括节点距离计算、关系距离计算、内部和外部 EMD 距离的计算. 求 EMD 距离可以归结为线性规划法, 在一般条件下其复杂度为多项式时间, 本文中分割的部分一般在 20 个以下 ( $n, m < 20$ ), 所以其复杂度为  $O(P(n, m))$ , 其中  $P(n, m)$  为由  $n$  和  $m$  组成的多项式. 其节点距离计算复杂度为一个定值, 因此其复杂度为  $O(1)$ ; 同理给定关系之间的距离计算复杂度也为  $O(1)$ . 计算任意两个节点间的内部 EMD 距离, 首先得到内部 EMD 矩阵(复杂度为  $O(n \times m)$ ), 再求出 EMD 距离(复杂度为  $O(P(n, m))$ ), 求出两个网页中所有节点之间的内部 EMD 距离(复杂度为  $O(P(n, m))$ )后再进行一次 EMD 计算便得到了外部 EMD(复杂度为  $O(P(n, m))$ ). 所以整个 NEMD 计算的时间复杂度为  $O(P(n, m))$ .

综合以上分析, 由于  $m, n$  远小于  $w_0$  和  $h_0$ , 因此整个算法的复杂度为  $O(n \times w_0 \times h_0)$ , 即本算法的复杂度主要在于图像分割处理部分.

## 5 性能分析

为了测试算法性能, 基于 matlab 实现了所提出的相似性检测算法和 Yu<sup>[6]</sup> 的相似性检测算法, 并对其进行了比较, 具体实验平台为普通 PC 机, CPU 为 P4 3.0, 内存 521MB, 操作系统为 Windows XP SP2.

首先测试图 4 中(a)和(b)之间的距离(进行了归一化处理, 0 为完全相同, 1 为完全不同), Yu<sup>[6]</sup> 的算法结果为  $1.107 \times 10^{-4}$  (表示二者十分相似), 本文算法的计算结果为 0.227 (表示不太相似), 可以看出本文算法反映了实际情况, 与用户的视图一致.

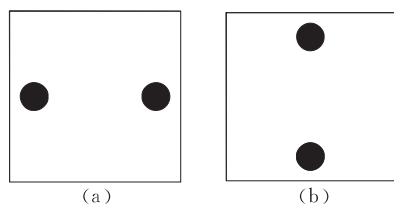


图 4 实验结果

同时我们还对实际的钓鱼网页进行了测试,钓鱼网页来自于 Liu<sup>[7]</sup> 的网站. 这些钓鱼网页针对 ebay 的有 2 个, 针对 earthlink, ICBC, Wells Fargo, US Bank 以及 Washington Mutual Bank 各一个. 同时 Liu 还提供了这 6 个网站的真实网页作为比较. 本文采用前缀“t-”来表示真实网页, 而“f-”表示钓鱼网页, 比如 t-eBay 表示真的 eBay, 而 f-ICBC 则表示针对 ICBC 的钓鱼网页. 表 1 和表 2 分别列出了根

据 Yu<sup>[6]</sup> 和本文方法算出的网页之间的距离. 可以看出绝大多数钓鱼网页与原网页都是最相似的(距离最小), 但由于针对 EarthLink 的钓鱼网页与原网页差距很大, 两个算法都出现错报(考虑到对用户的影响以及钓鱼网页的特点, 该错误是可以接受的, 因此在下面的计算中忽略该网页). 同时结果还显示了本文算法具有较好的鲁棒性.

表 1 Yu 的算法所计算出的网页距离

	t-eBay	t-Earth Link	t-ICBC	t-Wells Fargo	t-US Bank	t-Wash ington
f-eBay1	<b>0.0041</b>	<b>0.0292</b>	0.065	0.0432	0.0196	0.0256
f-eBay2	<b>0.0048</b>	0.0294	0.0643	0.0434	0.0203	0.0249
f-EarthLink	0.0187	<b>0.0293</b>	0.0609	0.0561	0.0248	0.0143
f-ICBC	0.0591	0.0633	<b>0.003</b>	0.0664	0.0566	0.0589
f-WellsFargo	0.0424	0.0571	0.0672	<b>0.0121</b>	0.0419	0.0559
f-US Bank	0.0172	<b>0.0240</b>	0.0596	0.0413	<b>0.0017</b>	0.0228
f-Washington	0.0293	<b>0.0231</b>	0.0597	0.0614	0.0299	<b>0.0095</b>

表 2 本文算法所计算出的网页距离

	t-eBay	t-Earth Link	t-ICBC	t-Wells Fargo	t-US Bank	t-Wash ington
f-eBay1	<b>0.0151</b>	0.2044	0.3483	0.1472	0.3458	0.2383
f-eBay2	<b>0.0032</b>	0.2051	0.3232	0.1452	0.3395	0.2405
f-EarthLink	0.1985	<b>0.1989</b>	0.4257	0.0820	0.3490	0.2449
f-ICBC	0.3219	0.4168	<b>0.0010</b>	0.4599	0.2155	0.4210
f-WellsFargo	0.1414	<b>0.1343</b>	0.4516	<b>0.0135</b>	0.2706	0.1685
f-US Bank	0.3370	0.3393	0.2153	0.2720	<b>0.0052</b>	0.3354
f-Washington	0.2470	0.2642	0.4280	0.1777	0.3387	<b>0.0125</b>

为了说明本文算法的鲁棒性, 我们对比两种算法所得的钓鱼网页与非钓鱼网页之间距离的比率. 设  $Sim(web_1, web_2)$  为网页  $web_1$  和  $web_2$  之间的距离, 则  $Sim(web_1, web_2)$  越小说明  $web_1$  和  $web_2$  越相似, 则  $Sim(t-web_i, f-web_j)$  与  $Sim(t-web_i, f-web_j)$  的比率就能反映出算法的准确性即相似度分辨率, 其中  $i \neq j$ , 由于有多个钓鱼网页, 我们采用最坏比率  $R_i^{worst}$  和平均比率  $R_i^{avg}$  进行比较, 如式(7)和(8)所示, 其中  $f-web_i^k$  为针对  $web_i$  的第  $k$  个钓鱼网页且  $i \neq j$ .

$$R_i^{worst} = \frac{\max(Sim(t-web_i, f-web_j))}{\min(Sim(t-web_i, f-web_i^k))}$$

(7)

$$R_i^{avg} = \frac{avg(Sim(t-web_i, f-web_j))}{avg(Sim(t-web_i, f-web_i^k))}$$

(8)

图 5 反映了两算法的最坏距离比率, 图 6 反映了两算法的平均距离比率, 可以看出本文算法具有较好的鲁棒性和精确性.

6 小 结

本文针对网络钓鱼发展和演变, 通过分析现有的主要防范措施, 提出了一套完整的钓鱼网页检测

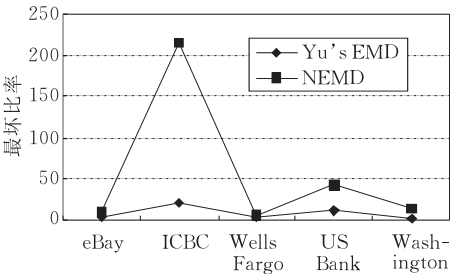


图 5 最坏比率

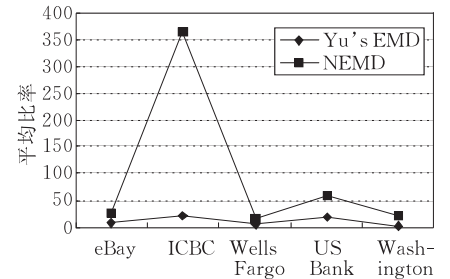


图 6 平均比率

体系架构, 基于该架构能实现钓鱼网站的发现、预警、取证等一系列完整的网络钓鱼防范措施. 深入研究了钓鱼网页检测中最为核心的算法——网页相似性计算, 提出了一个基于图像分割和嵌套 EMD 距离的网页相似性计算方法, 并通过实验证明了该方

法在准确性和鲁棒性方面优于已有的网页相似性检测算法。

下一步的工作是在一定范围内部署本系统,实现对真实网络钓鱼的检测,同时针对算法特点优化和改进图像分割算法以提高效率。

## 参 考 文 献

- [1] Dhamija Rachna, Tygar J D. The battle against phishing: Dynamic security skins//Proceedings of the 2005 Symposium on Usable Privacy and Security Table of Contents. Pittsburgh, Pennsylvania, 2005: 77-88
- [2] Inomata A, Rahman M, Okamoto T, Okamoto E. A novel mail filtering method against phishing//Proceedings of the Conference on Communications, Computers and signal Processing (PACRIM 2005), 2005: 221-224
- [3] Madhusudhanan Chandrasekaran, Ramkumar Chinchani, Shambhu Upadhyaya. PHONEY: Mimicking user response to detect phishing attacks//Proceedings of the 2006 International Symposium on World of Wireless, Mobile and Multimedia Networks Table of Contents. Washington, DC, USA: IEEE Computer Society, 2006: 668-672
- [4] Choi Daeseon, Jin Seunghun, Yoon Hyunsoo. A method for preventing the leakage of the personal information on the Internet//Proceedings of the 8th International Conference, Ad-

- vanced Communication Technology. Korea, 2006, 2: 20-22
- [5] Pan Ying, Ding Xuhua. Anomaly based Web phishing page detection//Proceedings of the 22nd Annual Computer Security Applications Conference. Washington, DC, USA, 2006: 381-393
- [6] Fu Anthony Y, L W, Deng Xiaotie. Detecting phishing Web pages with visual similarity assessment based on earth mover's distance (EMD). IEEE Transactions on Dependable and Secure Computing, 2006, 3(4): 301-311
- [7] Liu W, G H, Liu X, Zhang M, Deng X. Phishing Webpage detection//Proceedings of the 8th International Conference on Documents Analysis and Recognition. Seoul, Korea, 2005: 560-564
- [8] Nesbitt K V, Friedrich C. Applying gestalt principles to animated visualizations of network data//Proceedings of the 6th International Conference on Information Visualisation. Boston, USA, 2002: 737-743
- [9] Kim Duck Hoon, Yun Il Dong, Lee Sang Uk. A new attributed relational graph matching algorithm using the nested structure of earth mover's distance//Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004). Cambridge, UK, 2004, 1: 48-51
- [10] Zhang Heng-Bo, Ou Zong-Ying. An image search method based color and gray histograms. Computer Engineering, 2004, 30(10): 20-22(in Chinese)  
(张恒博, 欧宗瑛. 一种基于色彩和灰度直方图的图像检索方法. 计算机工程, 2004, 30(10): 20-22)



**CAO Jiu-Xin**, born in 1967, Ph.D., associate professor. His research interests include network security, service computing and digital right management (DRM).

**MAO Bo**, born in 1982, Ph.D. candidate. His research interests are network security with focus on web phishing

and digital city.

**LUO Jun-Zhou**, born in 1960, Ph.D., professor, Ph.D. supervisor. His current research interests include next generation network architecture, protocol engineering, network management and security, grid computing.

**LIU Bo**, born in 1975, Ph.D., lecturer. Her current research interests include pervasive computing and service management.

## Background

This research is supported by the National Natural Science Foundation of China under grant No.90604004 and No.60773103, China Specialized Research Fund for the Doctoral Program of Higher Education under grant No.200802860031, Jiangsu Provincial Natural Science Foundation of China under grant No.BK2007708 and No.BK2008030, Jiangsu Provincial Key Laboratory of Network and Information Security under grant No.BM2003201, Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education under grant No.93K-9.

Web Phishing is a new threat to web application, this problem has focused more and more attention from both industry and academic research since its impact to security and privacy negatively impairs Internet commerce especially in

online finance transactions. Phishing detection is an important tool to protect security in this area. In this paper, a novel architecture of Phishing web detection, and a visual based web page similarity detecting algorithm are proposed. The architecture outline the integrate function modules and processing workflow. The algorithm is composed of two novel algorithms, Web-image segmentation algorithm and Web matching algorithm. The Web-image segmentation algorithm is based on iterated dividing and shrinking, and Web matching algorithm based on Nested-EMD. According to the authors' experiments, the approach is better than other schemes in both accuracy and robustness. For future work, improvements can be done in feature extraction and relation construction. Some new matching rules also can be implemented, for example, by giving different key-zones to different weights.