Journal of Southeast University (English Edition) Vol. 26, No. 1, pp. 43 - 47 Mar. 2010 ISSN 1003—7985

Architecture and algorithm for web phishing detection

Cao Jiuxin[1] Wang Tianfeng[1] Shi Lili[1] Mao Bo[2]

([1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China)
([2]School of Architecture and Built Environment, Royal Institute of Technology, Stockholm SE-10044, Sweden)

Abstract: A phishing detection system, which comprises client-side filtering plug-in, analysis center and protected sites, is proposed. An image-based similarity detection algorithm is conceived to calculate the similarity of two web pages. The web pages are first converted into images, and then divided into sub-images with iterated dividing and shrinking. After that, the attributes of sub-images including color histograms, gray histograms and size parameters are computed to construct the attributed relational graph(ARG) of each page. In order to match two ARGs, the inner earth mover's distances (EMD) between every two nodes coming from each ARG respectively are first computed, and then the similarity of web pages by the outer EMD between two ARGs is worked out to detect phishing web pages. The experimental results show that the proposed architecture and algorithm has good robustness along with scalability, and can effectively detect phishing.

Key words: phishing detection; image similarity; attributed relational graph; inner EMD; outer EMD

Web phishing tricks users for their private information, (e. g., bank accounts, passwords, and credit card numbers) with phishing web pages which imitate the real web sites. According to the statistics from the anti-phishing work group (APWG) in the second half of 2008[1], the number of web phishing pages has maintained a high level, and the financial industry continues being the main target while the attacks on payment services are growing significantly. People are widely infected by web phishers, as a sign of which the crimeware-spreading sites infecting PCs with password-stealing crimeware has had a startling 827% increase from the beginning of 2008. The phishing problem has received much attention from both industry and academic research since its impact on security and privacy impairs Internet commerce especially on online financial transactions. Therefore, phishing detection is an important approach in guaranteeing the security of online services.

The existing anti-phishing schemes can be grouped into four categories: server-based, browser-based, server-browser cooperation and third-party-based.

Server-based schemes refer to those which require users' authentication to defend against phishing attacks. One famous example is the web site of the Bank of America, which asks users to select a personal image when registering and displaying the user-selected image with any forms that request a password. Users enter their password only when they confirm the original image they selected. However, Harvard and MIT researchers found that most online banking customers do not notice even if the site key images are not present, which indicates this method is not always effective[2].

Browser-based schemes embed anti-phishing measures into web browsers. This requires the browsers to maintain a list of known phishing sites and check the sites based on the list. Popular browsers such as IE 7/8, Firefox2, and Opera9.1 have already contained toolbars or plug-ins. The deficiency of this method, as can be easily deduced, is the time delay in real-time response. While a new phishing site is created, the browsers have to wait until some victims (possibly themselves) report it and the blacklist is updated; however, it is a fairly simple scheme and is easily deployed.

Server-browser cooperation, just as the name implies, calls for the teamwork of the server and the browser. Dynamic security skins[3-4], which requires users to recognize a photographic image, is a representative instance. The remote server generates an abstract image which is unique for each user and transaction. This image is used to create a "skin", which customizes the appearance of the server's web page. The browser computes the image it expects to receive from the server and displays it in the user's trusted window. In order to authenticate content from the server, the user can visually verify the images.

Third-party-based architectures, as the name implies, need an independent extra server which either filters the URL of the phishing sites or compare the fishing web pages with the legitimate ones. Email detection[5], network action testing[6], personal information protection[7], and visual similarity detection are some typical methods. Among them, visual similarity detection is a more efficient approach to identifying phishing webs.

There are two main approaches to web similarity detection, HTML and image-based solutions. Because of the flexibility of HTML and the diversity and dynamic configuration of web elements, it is not difficult to create two pages which are visually similar but completely different in HTML. This dynamic feature makes a HTML-based detection scheme ineffective under some circumstances. Larger scale experiments[8] show that the image-based approach is a better solution in both accuracy and robustness.

1 Related Work

Fu et al.[9] proposed a pixel and its location EMD based on the matching algorithm. With large-scale experiments, they demonstrated that their image-based solution had a better performance than HTML-based solutions. However, Fu's scheme only considers the pixel's absolute distribution rather than the relative distribution, which plays an important role in viewing according to the Gestaltian principle. If

Received 2009-07-31.

Biography: Cao Jiuxin (1967—), male, doctor, associate professor, jx.cao@seu.edu.cn.

Foundation items: The National Basic Research Program of China (973 Program) (2010CB328104, 2009CB320501), the National Natural Science Foundation of China (No. 60773103, 90912002), Specialized Research Fund for the Doctoral Program of Higher Education (No. 200802860031), Key Laboratory of Computer Network and Information Integration of Ministry of Education of China (No. 93K-9).

Citation: Cao Jiuxin, Wang Tianfeng, Shi Lili, et al. Architecture and algorithm for web phishing detection[J]. Journal of Southeast University (English Edition), 2010, 26(1): 43 - 47.

the related locations of blocks are changed, the scheme cannot detect the dissimilarity in some circumstances.

Cordero et al. [10] proposed another image-based detection which relied on images of rendered web pages to identify phishing attacks. However, it is difficult to find a good kernel for the support vector machine(SVM) [11] because of the mathematical sophistication it requires and it does not consider the partial match based on the sub-images of the web pages and their relative locations.

Pan et al. [12] proposed a web anomaly detection approach. However, Pan's document-object-model (DOM)-based phishing detection is not reliable as Fu's example shows[9] and cannot deal with pages containing lots of images. Because the image content is not a part of the DOM, the validity of the proposed rules in Pan's paper cannot be proved by real examples, and the SVM used in the approach needs an effective training process and it is difficult to implement.

For image-based detections, the basic step is to convert web pages to images. The projection profile cutting (PPC) [13] is an effective algorithm dealing with document images. However, web pages are so rich and colorful that the PPC is not strong enough to handle them. To efficiently divide one web image into sub-images, we propose a new web-image segmentation algorithm named the iterated dividing and shrinking algorithm. The web page is first transformed into an image, and then by shrinking and splitting repeatedly, the image is divided into sub-images.

In this paper, we put forward the architecture for detecting phishing pages. As a core part of the architecture, a new web matching algorithm based on the nested-EMD[14] is proposed to calculate the similarity of two web pages, in which the features are from the sub-images and the position relation vectors are added to calculate the inner distance of two ARGs. Furthermore, a system is recommended to implement the algorithm.

2 Architecture

The overall architecture described in Fig. 1 comprises three parts: the client browser, the anti-phishing center, and the protected server. The client browser has an anti-phishing plug-in which maintains a blacklist of phishing URLs or IP addresses, and a whitelist of the protected URLs or IP addresses. The plug-in has two functions. One is to filter the URLs based on the blacklist, and the other is to detect the sensitive information of a user's input in time. The anti-phishing center carries the analyzing work and appears as a distribution system. The suspected web pages captured by the browser plug-in are first delivered to the control center, and then are broadcasted to all the sub-centers. Each sub-center analyses the similarity solution of the suspected pages and the protected ones. The protected servers are the real web sites of the protected enterprises. The three parts work together cooperatively.

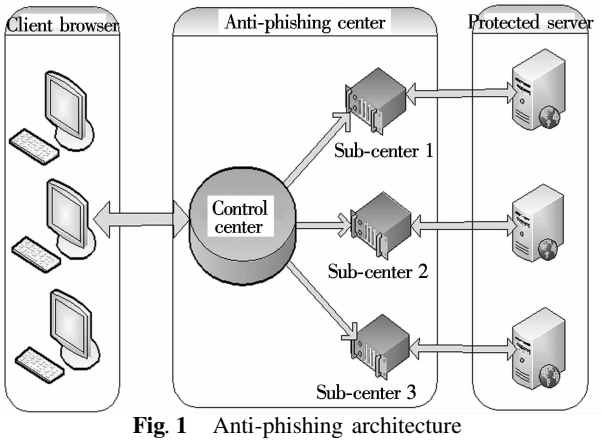

Fig. 1 Anti-phishing architecture

Fig. 2 shows the interaction sequence of the anti-phishing system. First, the blacklist and whitelist on the browser are initialized by the analysis center. Once the user's sensitive information is required, the pages in the blacklist are rejected directly by the plug-in; those in the whitelist are allowed, and others will be marked as suspected pages delivered to the analysis center. If the page appears to be similar to some legitimate one protected by some sub center, its URL and IP information will be sent to the analysis center, and the center will update the entire blacklists embedded in the client browser plug-ins.

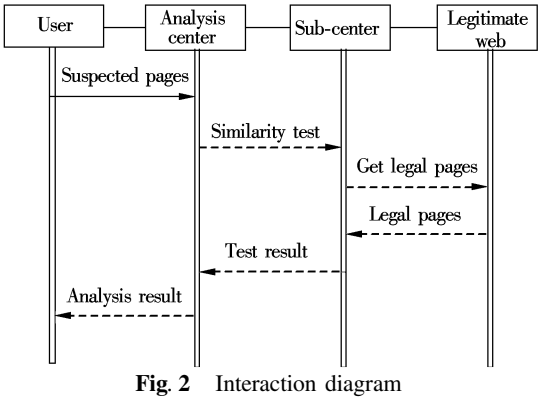

Fig. 2 Interaction diagram

Our approach focuses on the similarity detection algorithm which includes the detection method and similarity matching. First, convert the suspected web page into an image; next, segment the image into a set of sub-images that represent the blocks of the web page such as tables or frames; then, extract the features of each sub-image and compute the locational relationships among them. After that the web page is converted into an ARG with nodes representing key-zones and edges which imply their positional relationships. Finally, the nested-EMD is employed to compute the distance between two ARGs that are extracted from the suspected page and the real page. Based on the nested-EMD, the similarities of the two pages can be figured out, and the phishing judgment can be concluded.

3 Web Page Image Segmentation and ARG Generation

In the fields of computer vision and pattern recognition, the ARG graph is often used to express the object characteristics. One normal form of the ARG is

$$\boldsymbol{G} = \{\boldsymbol{V},\ \boldsymbol{R}\}$$
$$\boldsymbol{V} = \{a_i \mid 1 \leqslant i \leqslant n\};\ \boldsymbol{R} = \{\boldsymbol{r}_{ij} \mid 1 \leqslant i \leqslant n,\ 1 \leqslant j \leqslant n\}$$

where $\boldsymbol{G}$ is an ARG; $\boldsymbol{V}$ is the node set; a node presents a part of an object and its features; a_i presents the i-th node; n is the node number; $\boldsymbol{R}$ and $\boldsymbol{r}_{ij}$ denote the correlation set and the relationship between node a_i and node a_j, respectively. So the similarity match problem of two objects comes down to the match of two ARGs. In this paper, in

order to detect the similarities of web pages, every page is first converted into a page image, and then divided into sub-images. To construct the ARG in which each node corresponds to a divided sub-image, sub-image characteristics including color histograms(H), gray histograms(G) and size parameters(size) are computed. Then the attribute set of a sub-image can be expressed as $a = \{H, G, \text{size}\}$. All attribute sets compose a set vector $\boldsymbol{V} = \{a_i \mid 1 < i < n\}$ in which n is the number of sub-images and a_i is the attribute set of the i-th sub-image.

With the results from segmentation, we can obtain the coordinates of each sub-image, and then build a relation matrix which represents the relative position relations of all the sub-images in a web image. Every rectangle splits the plane into nine zones as shown in Fig. 3, and the position relation between A and any other sub-image such as B can be represented by a nine-dimensional vector $\boldsymbol{r}_{ij} = \{\langle z_1, z_2, \ldots, z_k, \ldots, z_9\rangle \mid z_k \in \{0, 1\}\}$ in which z_k denotes if the j-th sub-image is in the k-th zone of the i-th block, and 1 for yes, 0 for not. As shown in Fig. 3, the relationship between A and B is $r(A, B) = \{0, 0, 0, 1, 0, 0, 1, 0, 0\}$. Generally $r(A, B) \neq r(B, A)$, because if A is on B's left, B should be on A's right.

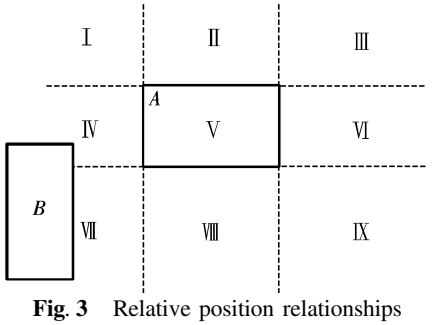

Fig. 3 Relative position relationships

All the relation vectors among the sub-images compose the relationship matrix $\boldsymbol{R} = \{\boldsymbol{r}_{ij} \mid 1 < i < n, 1 < j < n\}$ in which n is the number of sub-images and $\boldsymbol{r}_{ij}$ is the relation vector between the i-th and the j-th sub-images.

With the matrix and attributes, the ARG of a web page $\boldsymbol{G} = \{\boldsymbol{V}, \boldsymbol{R}\}$ is formed as shown in Fig. 4. Then the distance between two web pages can be computed based on their ARGs.

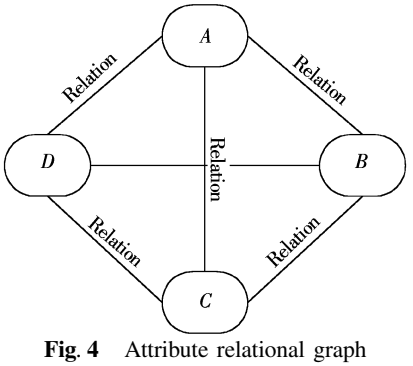

Fig. 4 Attribute relational graph

4 Similarity Computing with ARG Match

There are three steps in matching two ARGs. First, we obtain the feature and the relation distance. Then we compute the inner EMDs between every two nodes coming from each ARG respectively. Finally, we calculate the similarity of web pages by the outer EMD between two ARGs.

Assume that two ARGs of $\boldsymbol{G}$ and $\boldsymbol{G}'$ are matched, $\boldsymbol{G} = \{\boldsymbol{V}, \boldsymbol{R}\}$, $\boldsymbol{V} = \{a_i \mid 1 < i < n\}$, $\boldsymbol{R} = \{\boldsymbol{r}_{ij} \mid 1 < i < n, 1 < j < n\}$; and $\boldsymbol{G}' = \{\boldsymbol{V}', \boldsymbol{R}'\}$, $\boldsymbol{V}' = \{a_k' \mid 1 < k < n'\}$, $\boldsymbol{R}' = \{\boldsymbol{r}_{kl}' \mid 1 < k < n', 1 < l < n'\}$.

4.1 Feature distance

As shown in section 3 $a_i = \{H_i, G_i, \text{size}_i\}$, $a_{k'} = \{H_k, G_k, \text{size}_k\}$ and the distance between them $d(a_i, a_k') = \alpha S_{\text{H}} + \beta S_{\text{G}} + \gamma S_{\text{S}}$, where $\alpha + \beta + \gamma = 1$, S_{H}, S_{G}, and S_{S} are the similarities between the color histogram, the gray histogram, and the size, respectively.

1) Color and gray histogram similarity

Let H_p and H_q denote two color histograms, and the similarity S_{H} between them is

$$S_{\text{H}}(p, q) = \sum_{i=1}^{N} \min\{H_p(i), H_q(i)\}$$
$$\sum_{i=1}^{N} H_p(i) = \sum_{i=1}^{N} H_q(i) = 1; \quad N = 32 \qquad (1)$$

where $H_p(i)$ and $H_q(i)$ are the frequency of color i in color histogram H_p and H_q. The similarity between two gray histograms S_{G} is the same with S_{H},

$$S_{\text{G}}(p, q) = \sum_{i=1}^{N} \min\{G_p(i), G_q(i)\}$$
$$\sum_{i=1}^{N} G_p(i) = \sum_{i=1}^{N} G_q(i) = 1; \; N = 32 \qquad (2)$$

where $G_p(i)$ and $G_q(i)$ are the frequencies of gray i in the gray histograms G_p and G_q, respectively.

2) Size similarity

The size contains two components: width and height. Let $\text{size}_1 = \{w_1, h_1\}$, $\text{size}_2 = \{w_2, h_2\}$, and the size similarity S_{S} be

$$S_{\text{S}} = 1 - \frac{\min(w_1, w_2)\min(h_1, h_2)}{\max(w_1, w_2)\max(h_1, h_2)} \qquad (3)$$

where S_{S} is normalized and consistent with human perception.

4.2 Relation distance

The distance $d(r_{ij}, r'_{pq})$ between two nine-dimensional relation vectors can be computed by the EMD based on distance matrix $\boldsymbol{D}$. $\boldsymbol{D}$ is composed by the Manhattan distances of zones in Fig. 3. For example, the distance between zones 1 and 4 is 3, so $\boldsymbol{D}(1, 4) = 3$. By the matrix, we can obtain the EMD between two relation vectors. The relation distance matrix of Fig. 3 is

$$\boldsymbol{D} = \begin{bmatrix} 0 & 1 & 2 & 1 & 2 & 3 & 2 & 3 & 4 \\ 1 & 0 & 1 & 2 & 1 & 2 & 3 & 2 & 3 \\ 2 & 1 & 0 & 3 & 2 & 1 & 4 & 3 & 2 \\ 1 & 2 & 3 & 0 & 1 & 2 & 1 & 2 & 3 \\ 2 & 1 & 2 & 1 & 0 & 1 & 2 & 1 & 2 \\ 3 & 2 & 1 & 2 & 1 & 0 & 3 & 2 & 1 \\ 2 & 3 & 4 & 1 & 2 & 3 & 0 & 1 & 2 \\ 3 & 2 & 3 & 2 & 1 & 2 & 1 & 0 & 1 \\ 4 & 3 & 2 & 3 & 2 & 1 & 2 & 1 & 0 \end{bmatrix}$$

4.3 Inner EMD and outer EMD

The details of the inner EMD and the outer EMD are given by Kim[14]. The inner EMD indicates the difference of nodes in two ARGs, and the outer EMD indicates the correspondence of the two ARGs. By the distance of the NEMD, the similarity can be obtained for the two pages and is used to draw a conclusion if one page is a phishing page of the other.

Based on the feature and relation distances, the inner EMD between two nodes v_i and v_i' in G and G' can be calculated as follows.

First, we obtain the inner distance matrix of v_i and v_i', $\boldsymbol{D}_{\text{inner}} = [d_{\text{inner}}(j, j')]$. $d_{\text{inner}}(j, j')$ is computed by the following equation:

$$d_{\text{inner}}(j, j') = (1-p)\,d(v_j, v_{j'}') + pd(r_{ij}, r_{i'j'}') \quad (4)$$

where p is in the interval $[0, 1]$. Then, based on $\boldsymbol{D}_{\text{inner}}$, we obtain the inner EMD between nodes i and i'. This inner EMD yields one element of a distance matrix for the outer EMD.

Assume that a distance matrix for the outer EMD, $\boldsymbol{D}_{\text{outer}} = [d_{\text{outer}}(i, i')]$, is given from nn' inner EMDs. Then an outer EMD of G and G' can be computed with $\boldsymbol{D}_{\text{outer}}$. In order to allow for partial matches, all the weights in both the inner and the outer EMDs are identically provided as

$$w_i = w_{i'}' = \frac{1}{\max(n, n')} \qquad 1 \leqslant i \leqslant n;\ 1 \leqslant i' \leqslant n' \quad (5)$$

Algorithm 1 NEMD computing algorithm

```
/* computing the NEMD of ARGs: G, G' */
double getNEMD(G, G')
{
    double D_outer[n, n'] = 0; //outer matrix
    double S[n] = 1/max(n, n'); //attr. vector of page G
    double S'[n'] = 1/max(n, n'); //attr. vector of page G'
    for(int i = 0; i <= n; i++)
        D_outer[i, i'] = getInnerEMD(G, G', i, i');
    /* computing the outer EMD distance */
    double outEMD = EMD(D_outer);
    return outEMD;
}
/* computing the inner distance of i ∈ G and i' ∈ G' */
double getInnerEMD(G, G', i, i')
{
    double w = 0.5; //relation coefficient
    double D_inner[n, n'] = 0; //inner distance matrix
    /* computing inner distance matrix */
    for(int j = 0; j <= n; j++)
        D_inner[j, j'] = (1 - w) getDv(v_j, v_j') + wgetDr(r_ij, r'i'_j');
    /* computing the innerEMD */
    double innerEMD = EMD(D_inner);
    return innerEMD;
}
```

5 Experiments

In order to test the efficiency of our approach, we realized the web page dividing algorithm, the web similarity detection algorithm and Fu's algorithm[9]. The test data is supplied by Liu[8] in his homepage. Among Liu's data, two phishing web pages aimed at eBay and the ones at EarthLink, ICBC, Wells Fargo, US Bank, and Washington Mutual Bank, respectively. The correspondent six true target web pages are also collected for comparison. In the remaining part of this section, we denote a true webpage by adding the prefix "t-", e.g., t-eB stands for the true webpage of eBay, and a phishing webpage by adding the prefix "f-", e.g., "f-IC" refers to the phishing webpage targeted at ICBC.

Tabs. 1 and 2 show the results of Fu's EMD and our approach among the real and the phishing web pages, respectively. It is shown that both algorithms work well in the detection of the phishing page with the corresponding legitimate one except for EarthLink, because the phishing webpage itself of EarthLink is not similar to the real one in fact. According to the experience, almost all the phishing attackers devote their efforts to making their faked pages to be similar to the legitimate ones, and those non-similar ones are easy to be discovered by users. Therefore, examples such as EarthLink mentioned above are not under our consideration.

Tab. 1 Fu's EMD

EMD	t-eB	t-EL	t-IB	t-WF	t-USB	t-Wt
f-eB1	0.004 1	0.029 2	0.065	0.043 2	0.019 6	0.025 6
f-eB2	0.004 8	0.029 4	0.064 3	0.043 4	0.020 3	0.024 9
f-EL	0.018 7	0.029 3	0.060 9	0.056 1	0.024 8	0.014 3
f-IC	0.059 1	0.0633	0.003	0.066 4	0.056 6	0.058 9
f-WF	0.042 4	0.057 1	0.067 2	0.012 1	0.041 9	0.055 9
f-USB	0.017 2	0.024	0.059 6	0.041 3	0.001 7	0.022 8
f-Wt	0.029 3	0.023 1	0.059 7	0.061 4	0.029 9	0.009 5

Tab. 2 Nested EMD

EMD	t-eB	t-EL	t-IB	t-WF	t-USB	t-Wt
f-eB1	0.015 1	0.204 4	0.3483	0.1472	0.3458	0.2383
f-eB2	0.003 2	0.205 1	0.323 2	0.145 2	0.339 5	0.240 5
f-EL	0.198 5	0.198 9	0.425 7	0.082 0	0.349 0	0.244 9
f-IC	0.321 9	0.416 8	0.001 0	0.459 9	0.215 5	0.421 0
f-WF	0.141 4	0.134 3	0.451 6	0.013 5	0.270 6	0.168 5
f-USB	0.337 0	0.339 3	0.215 3	0.272 0	0.005 2	0.335 4
f-Wt	0.247 0	0.264 2	0.428 0	0.177 7	0.338 7	0.012 5

A metric, discriminative ratio (Dis$R$), is introduced to compare the robustness. Dis$R$ for each phishing site is expressed in Eq. (6), in which $\text{Sim}(\text{web}_1, \text{web}_2)$ is the similarity distance between web_1 and web_2. Dis$R$ presents the root-mean-square deviation based on the distance of the phishing page and its target.

$$\text{Dis}R_i = \sqrt{\frac{\sum_{j=0,\ j\neq i}^{n} [\text{Sim}(\text{tweb}_j, \text{fweb}_i) - \text{Sim}(\text{tweb}_i, \text{fweb}_i)]^2}{2}} \quad (6)$$

The results illustrated in Fig. 5 show that the Dis$R$ of the NEMD is much higher than that of Fu's EMD. It is shown that our scheme has a better performance in discrimination, and it detects phishing with strong robustness.

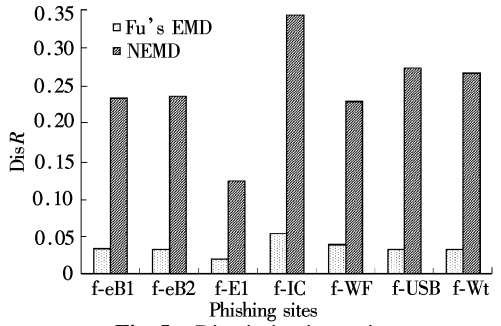

Fig. 5 Discriminative ratio

6 Conclusion

In this paper, an anti-phishing architecture and a novel phishing detection approach are proposed. The architecture comprises three parts: the client browser, the web server, and the analysis center. First, the web pages are converted into images, and then divided into sub-images; after that, the attributes of sub-images are computed to construct the ARG of each page. Finally, based on the ARG distance, the similarity of two web pages is calculated by the nested-EMD. Experimental results demonstrate that our approach is better than other schemes in both accuracy and robustness. Improvements can be made in feature extraction and relation construction. Some new matching rules can also be implemented by assigning different weights to different key-zones.

References

[1] APWG. Phishing attack trends report, 2nd half/2008 [EB/OL]. (2009-03-17) [2009-05-01]. http://www.antiphishing.org/reports/apwg_report_H2_2008.pdf.

[2] Bank of America. Page of privacy & security [EB/OL]. (2009-04-01) [2009-05-01]. http://www.bankofamerica.com/privacy/index.cfm?template=sitekey.

[3] Dhamija R, Tygar J D. The battle against phishing: dynamic security skins [C]//*Proceedings of the Symposium on Usable Privacy and Security*. Pittsburgh, PA, USA, 2005: 77 - 88.

[4] Dhamija R, Tygar J D. Phish and hips: human interactive proofs to detect phishing attack [C]// *Second International Workshop, HIP* 2005. Bethlehem, PA, USA, 2005: 127 - 141.

[5] Inomata A, Rahman S, Okamoto T, et al. A novel mail filtering method against phishing [C]// *Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*. Victoria, BC, Canada, 2005: 221 - 224.

[6] Chandrasekaran M, Chinchani R, Upadhyaya S. PHONEY: mimicking user response to detect phishing attacks [C]// *Proceedings of the International Symposium on World of Wireless, Mobile and Multimedia Networks*. Buffalo-Niagara Falls, NY, USA, 2006: 668 - 672.

[7] Choi Daeseon, Jin Seunghun, Yoon Hyunsoo. A method for preventing the leakage of the personal information on the Internet [C]//*The* 8*th International Conference on Advanced Communication Technology*. Phoenix Park, Korea, 2006, **2**: 1194 - 1198.

[8] Liu Wenyin, Huang Guanglin, Liu Xiaoyue, et al. Phishing web page detection [C]//*Proceedings of the Eighth International Conference on Document Analysis and Recognition*. Seoul, Korea, 2005, **2**: 560 - 564.

[9] Fu Anthony Y, Liu Wenyin, Deng Xiaotie. Detecting phishing web pages with visual similarity assessment based on earth mover's distance(EMD) [J]. *IEEE Trans on Dependable and Secure Computing*, 2006, **3**(4): 301 - 311.

[10] Cordero A, Blain T. Catching phish: detecting phishing attacks from rendered website images [R]. Berkeley, CA: University of California, 2006.

[11] Cortes C, Vapnik V. Support-vector networks [J]. *Machine Learning*, 1995, **20**(3): 273 - 297.

[12] Pan Ying, Ding Xuhua. Anomaly based web phishing page detection[C]//*The* 22*nd Annual Computer Security Applications Conference*. Miami Beach, FL, USA, 2006: 381 - 392.

[13] Nagy G, Seth S, Stoddard S D. Document analysis with an expert system [C]//*Proceedings of the* 7*th International Conference on Pattern Recognition in Practice*. Paris, France, 1986: 19 - 21.

[14] 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[C]//*Proceedings of the* 17*th International Conference on Pattern Recognition*. Cambridge, UK, 2004: 48 - 51.

一种网络钓鱼检测的体系结构及算法

曹玖新[1] 王田峰[1] 时莉莉[1] 毛 波[2]

([1] 东南大学计算机科学与工程学院, 南京 210096)

([2] 瑞典皇家理工大学结构与建筑环境学院, 斯德哥尔摩 SE-10044)

摘要: 提出了一个网络钓鱼防范系统, 该系统由客户端过滤插件、后台分析中心和受保护网站 3 个逻辑组件构成. 设计了一个基于图像的网页相似度检测算法, 该算法首先将被检测网页转换为图像格式, 然后采用迭代分割和收缩算法将原始图像划分为一组子图像集合, 在计算子图像颜色直方图、灰度直方图以及大小参数的基础上, 构建被检测网页的特征关系图(ARG),计算 ARG 之间的内部 EMD 距离, 并通过计算 2 个网页 ARG 之间的外部 EMD 距离来标示网页之间的相似度, 最终通过对不同网页之间相似度的分析检测出钓鱼网站. 实验结果显示所提出的体系结构与算法具有良好的鲁棒性和可扩展性, 可对钓鱼网页进行更加有效的检测.

关键词: 钓鱼检测; 图像相似度; 特征关系图; 内部 EMD; 外部 EMD

中图分类号: TP393