EMD based Visual Similarity for Detection of Phishing Webpages

Yingjie Fu, Liu Wenyin, Xiaotie Deng
Dept. of Computer Science, City University of Hong Kong, Tat Chee Avenue, Hong Kong, China
{anthony@cs., csliuwy@, csdeng@}cityu.edu.hk

Abstract

Phishing has become a severe problem in the Internet society. We propose an effective phishing webpage detection approach using EMD (Earth Mover's Distance) based visual similarity of webpages. Both suspected webpage and protected webpage are first preprocessed into low resolution images respectively. The image level colors and coordinate features are used to represent the image signatures. We then use the EMD method to calculate the signature distances of the two images as their visual similarity. When the visual similarity value is higher than a threshold, we classify the suspected webpage as a phishing webpage to the protected one. As our approach is based on image level color and coordinate features other than HTML, webpage obfuscation scams are neatly cracked. Large scale experiments with 1011 training webpages and 10,279 evaluation webpages are carried out to show high classification precision, phishing recall and applicable time performance for online enterprise solution.

1. Introduction

Phishing webpages are forged webpages created and used by phishers to mimic the webpages of certain real companies such that to spoof end users to leak their private information. Most of these kinds of webpages have high visual similarities with their targets such that their victims can trust them. Some of this kind of webpages even looks exactly the same as the real ones. Unwary internet users who access phishing webpages may be easily deceived by this kind of scams. Victims of phishing webpages may expose their bank account, password, credit card number, or other important information to the phishing webpage owners (phishers).

More and more phishing webpages have been found in recent several years. It has drawn high attention in both industry and academic research colonies. Report from Anti-Phishing Working Group [1] shows that the number of phishing attacks is increasing 50% for each month and usually 5% of the phishing email receivers will response to the scams.

In this paper, we propose an effective approach to detection of phishing webpages, which employs the Earth Mover's Distance (EMD) method [2] to calculate the visual similarity of webpages. Phishers are becoming cleverer to visually mimic the real webpages with various methods (e.g., images, flashes, and scripts etc.) rather than pure HTML only. We follow the anti-phishing strategy in [7] and [8] to obtain suspected webpages, which are first converted into normalized images and then represented with features composed of dominant color category and its corresponding centroid coordinate. The linear programming algorithm [11] for EMD is applied to calculate the visual similarity of two webpages based on their features. If the similarity exceeds the threshold associated to a protected webpage, we classify the webpage as a phishing webpage.

We use 1,011 training webpages (including 11 phishing webpages targeting at 8 true webpages), and 10,279 test webpages (include the same 11 phishing webpages) in our experiments. The results show high classification precision (99.93%), high phishing recall (90.91%) and satisfactory time performance (less than 0.1 second for one pair of similarity calculation).

The rest of this paper is organized as following. In Section 2, related work is introduced. In section 3, webpage preprocessing method and webpage signature calculation method are addressed. In Section 4, we present the approach to calculate visual similarity based on EMD. In Section 5 we discuss our phishing classification method. In Section 6, we present the experiments and results. Finally, we conclude our work and discuss future work in Section 7.

2. Related Work

The phishing problem has emerged for several years. Strong authentication of webpages [3] is widely used in security demanded websites. Commercial legislation actions against internet frauds are done in different countries. However, the most effective strategy for phishing detection is probably an active approach based on visual comparison, such as the one proposed by Liu et al. in [7] [8], which uses the region based approach to visual similarity of webpages.

The most straightforward way for a phisher to spoof people is to make their phishing webpages similar to the real ones. Previous research works provide various duplicate document detection approaches, which focus on plain text documents and use pure text features in similarity measure. In [7] [8], the visual similarity of webpages is oriented, and the concept of visual comparison based phishing detection was first introduced. It is an active anti-phishing approach that a phishing webpage can be detected and reported in an automatic way rather than involving too many human efforts. Their method first decomposes the webpages (in HTML) into salient (visually distinguishable) block regions. The visual similarity between two webpages is then evaluated in three metrics: block level similarity, layout similarity, and overall style similarity, all are based on matching of the salient block regions. The approach in this paper follows their overall strategy but use a different way to calculate visual similarity of webpages. We first convert HTML webpages into images and then employ the EMD method to the signatures of the images for similarity calculation.

EMD [2] is a method to evaluate the distance (dissimilarity) between two signatures. A signature is a set of features and their corresponding weights. The method comes from the well known transportation problem. Suppose we have *m* producers, each producer comes with a weight representing the amount of product he has. We denote producer set *P* as:

$$P = \{(p_1, w_{p_1}), (p_2, w_{p_2}), ..., (p_m, w_{p_m})\}$$

Suppose we also have *n* customers, each consumer comes with a weight indicating the amount of product he needs. We denote the consumer set *C* as:

$$C = \{(c_1, w_{c_1}), (c_2, w_{c_2}), ..., (c_n, w_{c_n})\}$$

Producers want to transport their products to consumers. Suppose the distances of each pair of producer and consumer are given, and they are represented into distance matrix *D*, which is defined before calculating EMD. It is represented as:

D=[d_{ij}], where $1 \le i \le m$ and $1 \le j \le n$

Producers produce the same product and consumers consume the same products. The transportation fee is proportional to both distance and product weight. The task is to find a flow matrix *F*, which contains factors indicating the amount of product to be moved from one producer to one consumer.

F=[f_{ij}], where $1 \le i \le m$ and $1 \le j \le n$

The transported product amount from *P* to *C* should be as much as possible and the total transportation fee should be minimized. The total cost of transportation fee can be represented as:

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} \cdot d_{ij} \tag{1}$$

The calculation of *F* is subject to the following constraints:

$$s.t.\begin{cases} f_{ij} \ge 0 \quad where\ 1 \le i \le m, 1 \le j \le n \\ \sum_{j=1}^{n} f_{ij} \le w_{pi} \quad where\ 1 \le i \le m \\ \sum_{i=1}^{m} f_{ij} \le w_{cj} \quad where\ 1 \le j \le n \\ \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = Min(\sum_{i=1}^{m} w_{pi}, \sum_{j=1}^{n} w_{cj}) \end{cases} \tag{2}$$

It is a Linear Programming (LP) problem. We solve it to get *F*, and then calculate EMD. The EMD can be represented as:

$$EMD(P, C, D) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(f_{ij} \cdot d_{ij})}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} \tag{3}$$

It has been practically proved that EMD has advantages in representing problems involving multi-featured signatures. EMD allows for partial matches in a very natural way, and is especially fit for cognitive distance evaluation, as shown in [6]. People have successfully used it for vision problems.

3. Preprocessing and Signature Generation

In preprocessing, we first convert a webpage into an image of a normalized size (e.g. 100*100 pixels) using the Lanczos algorithm [5]. We then calculate its signature, which comprises features and their corresponding weights. A feature comprises a color and its distribution centroid in the image. The original color of each pixel is represented using the ARGB (alpha, red, green, and blue) system with 4 bytes (32 bits). However, this color space is too huge and should be degraded. Finally, based one experiments we select 8 (instead of the original 256) scales to represent each color component. The tuning process is shown in [9]. A degraded color is represented with a 4-tuple <*A*, *R*, *G*, *B*>.

The centroid C_{dc} of each degraded color *dc* is simply calculated as the average of the coordinates of all pixels with the same color. A feature F_{dc}, which has degraded color *dc* can be represented with *dc* and C_{dc}, F_{dc}=<*dc*, C_{dc}>. The weight corresponding to this feature is N_{dc}, which is total number of pixels with *dc*. So a complete signature *S* of an image is represented

as: $S = << F_{dc_1}, N_{dc_1} >, < F_{dc_2}, N_{dc_2} >, ..., < F_{dc_N}, N_{dc_N} >>$, where N is the total number of degraded colors. The feature-weight tuples in S are ranked in descending order of their weights, i.e., $N_{dc_i} \geq N_{dc_{i+1}}$ and $1 \leq i \leq N-1$. In our approach, we choose the first N_s feature-weight tuples of S to be the signature, and we denote it as S_s. Normally, N_s is less or equal to N. However, in case N is less than N_s, S is equal to S_s.

4. Computing Visual Similarity with EMD

We use the EMD method to calculate the visual similarity of webpages based on their image signatures. In the problem, the producer set is represented by one webpage's signature and the consumer set by the other.

The distance matrix $D=[d_{ij}]$ $(1 \leq i \leq m, 1 \leq j \leq n)$ of a pair of feature is calculated as follows. We calculate the normalized Euclidian distance ND_{color} of the two degraded ARGB colors and the normalized Euclidian distance $ND_{centroid}$ of their centroids, especially, and d_{ij} is simply their weighted average.

Given two features $\varphi_i = < dc_i, C_{dc_i} >$ and $\varphi_j = < dc_j, C_{dc_j} >$, where $dc_i = < dA_i, dR_i, dG_i, dB_i >$, and $dc_j = < dA_j, dR_j, dG_j, dB_j >$, the maximum color distance MD_{color} is defined as the length of the longest diagonal of the color space and the maximum centroid distance $MD_{centroid}$ is the length of the normalized image's diagonal, respectively. The normalized color distance ND_{color} is defined as:

$$ND_{color}(dc_i, dc_j) = \frac{\sqrt{(dc_i - dc_j) \times (dc_i - dc_j)^T}}{MD_{color}} \quad (4)$$

and the normalized centroid distance $ND_{centroid}$ is:

$$ND_{centroid}(C_{dc_i}, C_{dc_j}) = \frac{\sqrt{(C_{dc_i} - C_{dc_i}) \times (C_{dc_i} - C_{dc_i})^T}}{MD_{centroid}} \quad (5)$$

Finally, $d_{ij} = ND_{feature}(\varphi_i, \varphi_j)$ is just the weighted average of ND_{color} and $ND_{centroid}$.

Given two signatures $S_{s,a}$ and $S_{s,b}$, where $S_{s,a}$ has m features and $S_{s,b}$ has n features, the flow matrix $F_{ab}=[f_{ij}]$ $(1 \leq i \leq m, 1 \leq j \leq n)$ can be calculated through linear programming and then $EMD(S_{s,a}, S_{s,b}, D)$ between $S_{s,a}$ and $S_{s,b}$ can be calculated using the formula (3) in Section 2

Finally, we define the EMD based visual similarity $VS(S_{s,a}, S_{s,b})$ of two (webpage) images as:

$$VS(S_{s,a}, S_{s,b}) = 1 - [EMD(S_{s,a}, S_{s,b}, D)]^{\alpha} \quad (6)$$

where $\alpha \in (0, +\infty)$ is the amplifier of visual similarity to control the mapping between EMD and similarity.

5. Classification

Each protected webpage is associated with a unique threshold, which is used to determine whether it is attacked by a suspected webpage. If their similarity is larger than the threshold, the suspected webpage is considered as a phishing webpage to the protected webpage. We use different thresholds for different protected webpages because our defined similarity may not be well-proportioned in the feature space. For each protected webpage, we randomly collect some webpages from the Web and combine them with their reported phishing webpages (if there is any) and use them as the training dataset to determine the threshold. We select the threshold such that it can result the least wrong classifications, which can be either false positive or false negative.

6. Experiments

We randomly collected 10,268 homepages from the Web. In addition, we have 11 phishing webpages which targeted at 8 real webpages. The 10,268+11 webpages are mixed together to form the Suspected Webpage Set. The targeted 8 real webpages form the Protected Webpage Set. These datasets and the complete experiment results are available at our website [9].

As the parameter setting is very important, we tuned them based on a large scale of experiments. The detail of the tuning experiments is also available at [9]. Finally, we decide the parameters as follows: w=h=100, α=0.5, $|S_s|$=20 for the best classification performance.

From the 10,268 collected webpages, we select 1,000, which are used with 11 phishing webpages together as the training dataset to calculate the thresholds for the 8 protected webpages. The result thresholds are in Table 1.

Table 2 shows the classification precision, phishing recalls, and false alarms using these thresholds but on the entire Suspected Webpage Set. Most of the phishing webpages have been recognized except for two. One is "fake-ICBC(Asia)", which is missed. Instead, a very similar one, "www.frlp.utn.edu.ar", is found out. The reason of the wrong classification is that the real and phishing webpage of ICBC(Asia) are not visually similar to each other at all. The other one is "fake-Washington Mutual (2)", whose visual similarity ranks the 4th highest against "real-Washington Mutual".

Table 1. Thresholds for Protected Webpage Set (trained with 1000+11 suspected webpages)

Protected Webpage	Threshold
real-Bank of Oklahoma - Online	84.69
real-ebay1	94.34
real-eBay2	94.93
real-ICBC(Asia)	73.85
real-Key Bank	93.23
real-us bank	95.73
real-Washington Mutual	85.41
real-Wells Fargo Sign On	92.55

Table 2. Classification Precision, Phishing Recall, and False Alarm List (evaluated with 10,268+11 suspected webpages)

Protected Webpage	Classification Precision	Phishing Recall	False Alarm
real-Bank of Oklahoma	10279/10279	1/1	0
real-ebay1	10279/10279	3/3	0
real-eBay2	10279/10279	1/1	0
real-ICBC(Asia)	10275/10279	0/1	4
real-Key Bank	10279/10279	1/1	0
real-us bank	10279/10279	1/1	0
real-Washington Mutual	10278/10279	1/2	1
real-Wells Fargo	10279/10279	1/1	0
Overall	99.95%	81.82%	5

Table 3. Classification Precision, Phishing Recall, and False Alarm List (with looking forward 3 webpages)

Protected Webpage	Classification Precision	Phishing Recall	False Alarm
real-Bank of Oklahoma	10276/10279	1/1	3
real-ebay1	10276/10279	3/3	3
real-eBay2	10276/10279	1/1	3
real-ICBC(Asia)	10272/10279	0/1	7
real-KeyBank	10276/10279	1/1	3
real-us bank	10276/10279	1/1	3
real-Washington Mutual	10277/10279	2/2	2
real-Wells Fargo	10276/10279	1/1	3
Overall	99.74%	90.91%	27

In practical applications we are expecting to achieve better phishing recall even though the classification precision and false alarm could be sacrificed a little. We can look forward to report more phishing webpages to achieve better recall. Table 3 shows the classification precision, phishing recall, and false alarm values by looking forward 3 webpages after the corresponding thresholds.

Experiments also show that a pair of EMD based visual similarity calculation time is less than 0.1 second on an ordinary computer. It is fast enough for online anti-phishing task.

7. Conclusion and Future Work

We have proposed a method to classify phishing webpages from the suspected ones using the EMD based visual similarity on the assumption that phishing webpages should have similar appearance to the real ones. This method is image based rather than HTML based, by which phishing webpage obfuscation scams are cracked. Each webpage is preprocessed into an image signature, based on which the visual similarity is calculated using the EMD method. The method can be used in the anti-phishing strategy proposed in [7] and [8]. Experiments show that our method can achieve satisfying classification precision and phishing recall. The time efficiency of computation is also acceptable for online phishing detection. In future work, we will continue with more phishing examples and even larger scale datasets. We will also fully compare its performance with the methods proposed in [7] and [8].

Acknowledgement

The work described in this paper was fully supported by grants from City University of Hong Kong (Project No. 7001462 and 7001545).

References

[1]. Anti-Phishing Working Group, http://www.antiphishing.org.

[2]. Hitchcock F. L. *The distribution of a product from several sources to numerous localities*. J. Math. Phys., 20:224-230, 1941.

[3]. http://wp.netscape.com/eng/ssl3/

[4]. Jakobsson M. *Modeling and Preventing Phishing Attacks*, Phishing Panel of Financial Cryptography, 2005

[5]. John C. R., *The image processing handbook second edition*, CRC Press, 1995

[6]. Levina E., Bickel P. *The Earth Mover's Distance is the Mallows Distance: Some Insights from Statistics*. ICCV2001,, Vol. 2

[7]. Liu W., Huang G., Liu X., Zhang M., Deng X., *Detection of Phishing Webpages based on Visual Similarity*, Poster, WWW2005, pp.1060-1061.

[8]. Liu W., Huang G., Liu X., Zhang M., Deng X., *Phishing Webpage Detection*, to appear in Proc. ICDAR 2005.

[9]. Fu A.Y., www.cs.cityu.edu.hk/~anthony/AntiPhishing

[10]. Wyszecki G. and Styles W.S., *Color Science: Concepts and Methods, Quantitative Data and Formulae*. John Wiley and Sons,New York, NY, 1982.

[11]. Hillier F. S., Liberman G. J. *Introduction to Mathematical Programming*. McGraw-Hill, 1990.