Detection of Phishing Webpages based on Visual Similarity

Liu Wenyin[1], Guanglin Huang[1], Liu Xiaoyue[2], Zhang Min[1,3], Xiaotie Deng[1]
[1]Department of Computer Science, [2]Department of Chinese, Translation & Linguistics
City University of Hong Kong, Tat Chee Avenue, Hong Kong SAR, PRC
[3] State Key lab of Intelligent Tech. & Sys., Tsinghua University, Beijing, 100084, PRC
(852) 2784 4730
{csliuwy@,hwanggl@cs.,xyliu0@, csdeng@}cityu.edu.hk, z-m@tsinghua.edu.cn

ABSTRACT
An approach to detection of phishing webpages based on visual similarity is proposed, which can be utilized as a part of an enterprise solution for anti-phishing. A legitimate webpage owner can use this approach to search the Web for suspicious webpages which are visually similar to the true webpage. A webpage is reported as a phishing suspect if the visual similarity is higher than its corresponding preset threshold. Preliminary experiments show that the approach can successfully detect those phishing webpages for online use.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval–Information filtering; I.7.5 [**Document and Text processing**]: Document Capture–Document analysis.

General Terms
Algorithms, Experimentation, Security, Human Factors

Keywords
Anti-Phishing, Web document analysis, Information filtering, Visual similarity

1. INTRODUCTION
Phishing [1] is a criminal trick of stealing victims' personal information by sending them spoofed emails urging them to visit a forged webpage that looks like a true one. The victims may finally suffer losses of money or other kinds.

We propose a method to detect the phishing webpages based on visual similarity. An important feature of a phishing webpage is its visual similarity to its target (true) webpage. Hence, a legitimate webpage owner or its agent can detect suspicious URLs and compare the corresponding webpages with the true one in visual aspects. If the visual similarity of a webpage to the true webpage is high, the owner will be alerted and can then take whatever actions to immediately prevent potential phishing attacks and hence protect its brand and reputation.

In our approach, the visual similarity between two webpages is measured in three metrics: block level similarity, layout similarity, and overall style similarity. All these three visual similarity metrics are defined based on webpage segmentation. A webpage is first decomposed into a set of salient blocks [2], [3]. The block level similarity is defined as the weighted average of the similarities of all pairs of matched blocks. The layout similarity is defined as the ratio of the weighted number of matched blocks to the number of total blocks in the true webpage. The overall style similarity is calculated based on the histogram of the style feature. The normalized correlation coefficient of the two webpages' histograms is the overall style similarity.

WWW 2005, May 10-14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

2. SYSTEM ARCHITECTURE
Figure 1 illustrates the system architecture of our approach. The true webpage is processed by the True Webpage Processing Module to obtain an intermediate representation and the visual features of the blocks. The Suspicious URL Detection Module generates certain suspicious URLs based on transformation of the true URL or detects the suspicious URLs from emails. For each webpage at a suspicious URL, the Suspicious Webpage Processing Module fetches the webpage at that suspicious URL if it is available and generates its representation. The Visual Similarity Assessment Module compares the true webpage and each suspicious webpage and calculates their visual similarity based on their intermediate representations. If the visual similarity between a suspicious webpage and the true one exceeds a threshold, the Phishing Report Module is called.

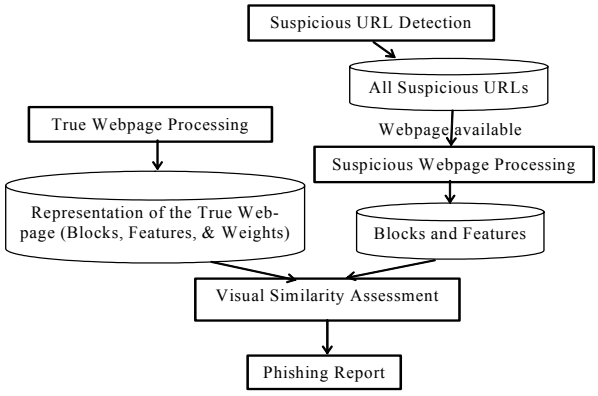

Figure 1. System architecture of the proposed approach

3. SIMILARITY ASSESSMENT
The Visual Similarity Assessment Module measures the visual similarity between two webpages in three aspects: block level similarity, layout similarity, and overall style similarity. All these three aspects are defined based on webpage segmentation. Next, we define these three similarity metrics in the following three subsections respectively.

3.1 Block Level Similarity

The block level similarity measures the visual similarity of two pages at the level of individual blocks. It is defined as the weighted average of the visual similarities of all matched block pairs between two pages. Basically, the content of a block can be categorized as either text or image. We use different features to represent text blocks and image blocks. The features for text blocks include colors, border style and alignment, etc., and the features for image blocks include alternative text, dominant color, and image size, etc.

We first calculate their similarity in terms of each feature in the feature set and then use a weighted sum of the individual feature similarities as the total similarity of the two blocks. The weight of each feature means its importance to the total similarity and can be assigned empirically. In our implementation, we focus more on color related features.

Two blocks are considered as matched if their similarity is higher than a threshold. After we obtain the similarity values of all pairs of possible matching blocks, we find a matching scheme between the two webpages' blocks. This is actually a bipartite graph matching problem and a globally optimal solution can be obtained.

3.2 Layout Similarity

Usually, it takes many efforts to make a brand new webpage mimicking a true webpage. A convenient way is to copy the source file of the true one and modify it a little bit for this purpose. In this case, the main webpage structure is kept. Hence, we define the layout similarity as the ratio of the weighted number of matched blocks to the total number of blocks in the true webpage. We employ the method in [3] for layout matching. Two blocks are considered matched if they both exhibit high visual similarity and satisfy the same constraints with corresponding matched blocks.

We then define the layout similarity of two webpages as the ratio of the weighted number of matched blocks to the total number of blocks in the true webpage, and the weight of each block is assigned differently according to its importance to the whole webpage.

3.3 Overall Style Similarity

In addition to the webpage content, the style consistency is another important feature which can easily cheat the victims' eyes. Generally, all webpages owned by one company would keep the style consistent. The overall style similarity focuses on the visual style of a webpage, which can be represented by several format definitions, e.g., the font family, background color, text alignment, and line spacing.

We first obtain the histogram of the style feature values for each webpage. For each feature value of one feature, we use the blocks and their weights as the unit to count their times, namely, the distribution value of that value. The overall style similarity of two webpages is calculated as the normalized correlation coefficient of the two webpages' histograms.

4. EXPERIMENTS

We have implemented our approach based on the three similarity metrics defined in Section 3 for experiments. We have collected 8 phishing webpages reported by [1] to evaluate our approach. These 8 phishing pages tried to attack 6 true webpages, which are also collected for comparison.

First we test our approach by comparing each phishing webpage with all true webpages in the test set. The result of our first trial indicates that, for most cases the real pairs of phishing webpages and their targets result in significantly higher similarity values than other pairs. This result shows that our similarity assessment metrics are suitably defined and compatible with human visual perception.

Furthermore, to test our approach's ability to avoid false alarms, we have also collected a set of 320 index pages from the official websites of 320 commercial banks. All these test webpages can be downloaded at [4]. The 6 true pages are regarded as query to search for visually similar webpages in the test dataset. We have done our experiment with the similarity threshold t=0.9 and t=0.7, respectively. The first setting results in no false alarm but one missing. The second setting results in 4 false alarms in total but no missing. Actually, we could adjust the threshold further to guarantee no missing but with less false alarms.

We also recorded the time cost for similarity calculation for each pair of pages and the average is around 0.8s. We think the speed of our approach should be fast enough for online detection of the phishing webpages, since the bottleneck for such application is usually at network latency.

5. CONCLUSIONS AND FUTURE WORKS

In this paper, we propose a novel approach to detecting phishing webpages based on visual similarity. The approach first decomposes the webpages into salient blocks according to visual cues. The visual similarity between two webpages is then measured in three aspects: block level similarity, layout similarity, and overall style similarity. A webpage is reported as a phishing suspect if any of these similarities to the true webpage is higher than a threshold.

We have done experiments on a test dataset of 328 suspicious webpages. Preliminary results show that our approach can successfully detect the phishing webpages with few false alarms for online use. We believe that the approach can be used as a part of an anti-phishing strategy in an enterprise solution. In future works, we plan to build a larger test dataset and thoroughly test this approach. We will also try to improve its efficiency and consider commercial application situations.

6. REFERENCES

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

[2] Chen Y., Ma W.Y., and Zhang H.J. Detecting webpage structure for adaptive viewing on small form factor devices. In *Proceedings of the 12th International Conference on World Wide Web*, pages 225–233, 2003.

[3] Liu Y., Liu W., and Jiang C. User interest detection on webpages for building personalized information agent. In *Proceedings of the Fifth International Conference on Web-Age Information Management (WAIM 2004)*, Dalian, China. LNCS, Vol. 3129, pages 280–287, 2004.

[4] http://www.cs.cityu.edu.hk/~liuwy/phishing/testdata.zip