ISSN:0975-9646

Phishing WebPages Strategy Based On Optical Semblance Assessment

Vijaykumar kangala, A. Ranjith Kumar ,Shilpa shesham ,Rampelly shilpa
CSE,Vivekananda Institute of Technology and Science SET, Karimnagar, AP, India

***Abstract*-An approach to phishing webpage strategy based on optical semblance assessment is proposed, which can be utilized as a part of an enterprise solution to anti-phishing. The approach first decomposes the webpages into salient (visually distinguishable) block regions. The optical semblance between two webpages is then evaluated in three metrics: block level similarity, layout similarity, and overall style similarity. A webpage is reported as a phishing suspect if any of them (with regards to the true one) is higher than its corresponding preset threshold. Preliminary experiments show that the approach can successfully detect those phishing webpages with few false alarms at a speed adequate for online application.**

Keywords: Phishing, Web document analysis, Information filtering, Optical semblance

1. INTRODUCTION

Phishing is a form of online identity theft that aims to steal sensitive information such as online banking password and credit card information from user. The victims may finally suffer loss of money or other kinds. We propose a method to detect the phishing webpages strategy based on optical semblance. An important feature of a phishing webpage is its optical semblance 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 optical semblance of a webpage to the true webpage is high, the owner will be alerted and can then take immediate actions to prevent potential phishing attacks and hence protect its brand and reputation. In this approach, the optical semblance between two web pages is measured in three metrics: Block Level Similarity, Layout Similarity, and Overall Style Similarity. All these three optical semblance metrics are defined based on webpage segmentation.

A webpage is first decomposed into a set of salient blocks. 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 weighted number of matched blocks to the number of total blocks in the true webpage. The overall style similarity is weighted average of the similarities of all pairs of matched blocks. Figure 1.1 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 E-Mails.

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.

2. OBJECTIVES

An approach to phishing WebPages strategy based on optical semblance assessment 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 optical semblance to the true webpage. A webpage is reported as a phishing suspect if the optical semblance is higher than its corresponding preset threshold.

The optical semblance Assessment Module compares the true webpage and each suspicious webpage and calculates their visual similarity based on their intermediate representations. If the optical semblance between a suspicious webpage and the true one exceeds a threshold, the Phishing Report Module is called.

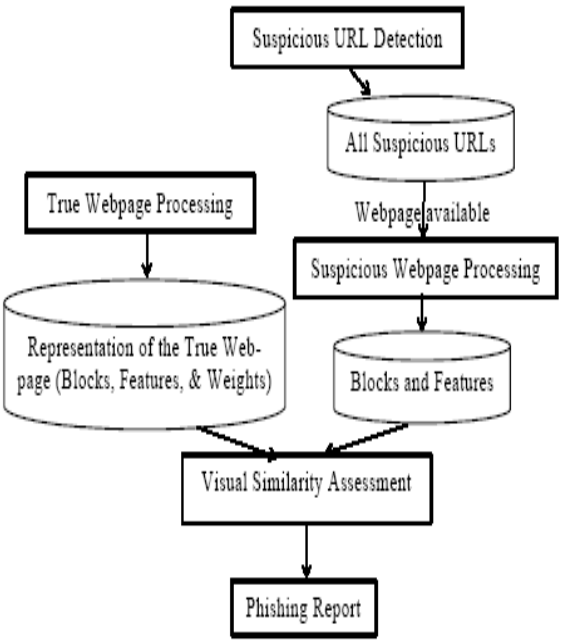

Figure 1.1: The System Architecture

2.1. Existing System

Phishing Web pages generally use similar page layouts, styles (font families, sizes, and so on), key regions, and blocks to mimic genuine pages in an effort to convince Internet users to divulge personal information, such as bank account numbers and passwords**.**

Phishers use a wide variety of technologies, with one common thread. All technologies employed by phishers have the goal of deception. It is not possible to determine whether a connection to a site is secure by looking at a **lock icon** in a browser. There are several reasons why a lock icon cannot be trusted:

❖ A lock icon by itself means only that the site has a certificate; it does not confirm that the certificate matches the URL being (deceptively) displayed. A user must click on a lock icon to determine what it means, and few ever do.

❖ It is possible to get a browser to display a lock icon using a self-signed certificate (i.e. a certificate that has not been issued by a valid certificate authority), with certain encryption settings.

❖ A lock icon may be overlaid on top of the browser using the same technologies used to fake the URL bar. This technology may even be used to present authentic-looking certificate data if the user clicks on the lock icon to confirm legitimacy.

To confront those challenges, we developed an **Antiphishing Strategy** that uses a visual approach to detect bogus webpages. To monitor phishing attacks, site owners can register their true URLs and associated keywords with our Site Watcher system. Given that most phishing attacks are initiated via E-Mail, Site Watcher is designed to run on mail servers and monitor and analyze both incoming and outgoing messages for potential phishing URLs.

2.2. Proposed System

When the website monitor deployed on a mail server identifies a message that contains a keyword requested by a customer, it sends the suspicious and true URLs to the optical semblance assessment module for further investigation.

This module extracts the Webpages features and measures the semblance to the true pages according to three metrics:

✓ Block-level (Detail),
✓ Layout (Global), and
✓ *Style* (Overall).

If the optical semblance is higher than the corresponding threshold, the system issues a phishing report to the customer.

We first decompose a page into a set of salient blocks — each visually (in terms of visual features) and semantically (in terms of content relevancy) consistent but distinguishable from adjacent blocks and providing granularity for use in analyzing the page's features.

At *block-level similarity*, we match blocks on the true Web page with the most similar blocks on the corresponding suspicious page. The system defines block-level similarity as the weighted average of the similarities of all matched-block pairs. At the *layout level similarity*, matching relies not only on individual block similarities but also on position constraints among blocks. We define layout similarity as the ratio of the weighted number of matched blocks to the number of total blocks in the true page. The system calculates *overall style similarity* based on a histogram of each page's style feature values (that is, the distribution of feature values). We use the correlation coefficient (normalized to the range of [0, 1]) of the two page's histograms as the overall style similarity.

3. METHODOLOGY:

Our project comprises the following modules which are performed the antiphishing and user interaction. The modules are:

1. User Module
2. Visual feature Extractor
3. Optical semblance module
   a. Block-level similarity
   b. Layout-level similarity
   c. Over all Style similarity
4. Phishing report.

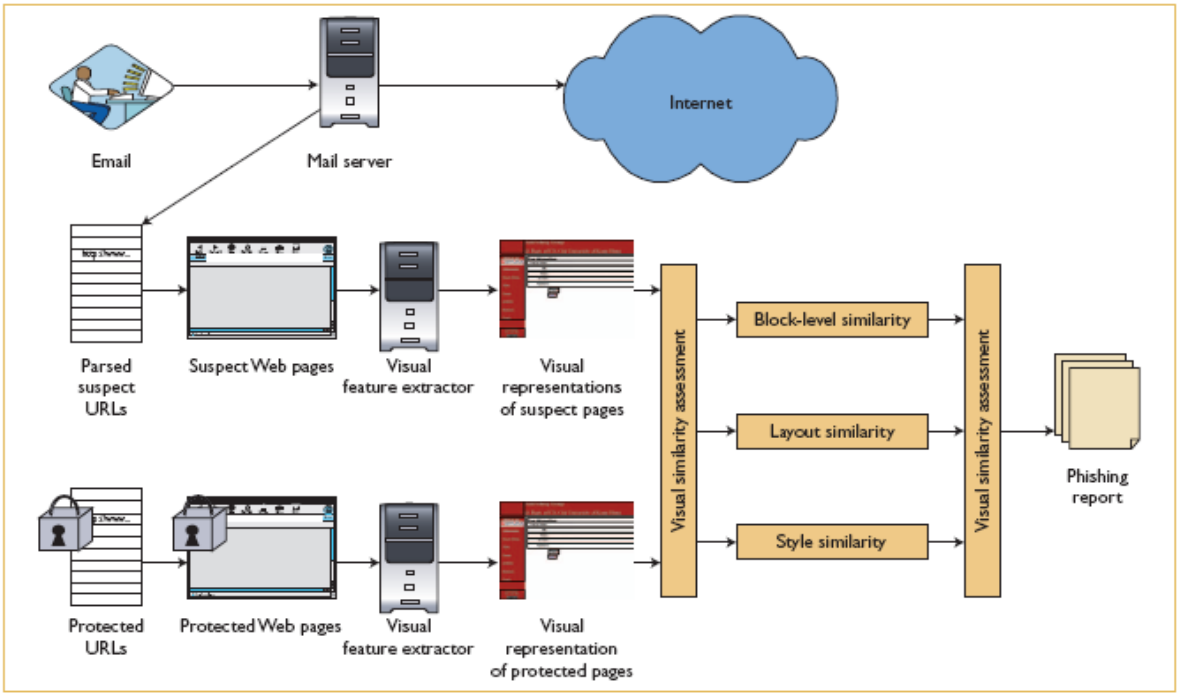

Figure 2.2.1 Block diagram that illustrates the system flow in this approach.

3.1. User Module: (Client Request/Web-Page Creation):
The URL given by the user is the input. The input URL given by the user is checked whether it has been registered or not. If the given URL is a registered one, then that URL is an authentic URL. Immediately the requested web page is displayed. If that page is not a registered one, then it undergoes a process of Visual Extraction.

3.2. Visual Feature Extractor
If the input URL given by the user is not a registered one, Visual Extraction process will take place. This is the process in which the True page and the suspicious page are separated. True page contains XML content and the suspicious page contains text content.

3.3. Optical semblance assessment
The three aspects that the optical semblance assessment module uses to compare pages are defined based on the salient blocks identified in the segmentation phase. This phase carried out three levels. They are:

a. Block-level similarity
b. Layout-level similarity
c. Over all Style similarity

3.3.1. Block-level similarity
We define block-level similarity as the weighted average of the optical semblance of all matched block pairs between two pages.
The module first categorizes block content as either text or image and then extracts the features from the blocks. Two block's total similarity is a weighted sum of the individual feature similarities. In our implementation, we focus more on color features, such as the block's background and foreground colors. If a feature's possible values are enumerative or discrete (for example, the font family can be Arial, Times, and so on), the similarity value for that feature is binary that is, 1 if the feature values are the same, and 0 otherwise. Two blocks are considered to match if their similarity average value is higher than a given threshold. After obtaining similarity values for all possible matching block pairs, it equals or higher then threshold add to list.

3.3.2. Layout similarity
In measuring layout similarity, we begin by finding several blocks with identical contents and then use the so-called *neighborhood relationship model* to match other blocks according to the spatial relations of all blocks on the page.
Now, consider two blocks to be matched if both exhibit high optical semblance and satisfy the *same position constraints* (layout relationships) *with corresponding already-matched blocks.*
If block 2 in true page and false page were matched, for example, then block 1 in true page should match block 1 in false page to score highly in layout similarity.

3. 3.3. Overall Style similarity
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.
If the styles are similar, most viewers have trouble telling whether a page is genuine or mimicked. Several format definitions that we use to represent webpages visual styles. We define overall style similarity between two pages as the correlation coefficient (normalized to the range of [0, 1]) of the pages.To compare similarity based on pure text features (keyword vectors), we also conducted *the experiments calculating similarity based on the pages TF/IDF (term frequency/inverse document frequency) weighted keyword vectors.*
The resulting similarity values were so close to establish a clear-cut threshold in the range. To successfully detect all phishing Web pages.

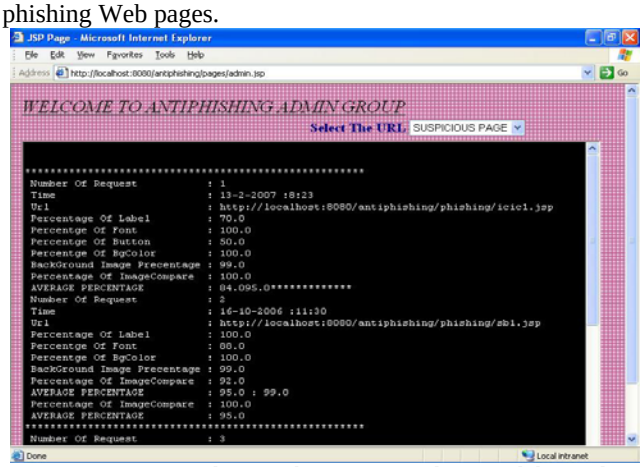

Figure 3.1: Report indication the various attributes and their values in the suspected URLs

3.4. Phishing Report
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. Suspect URL is retrieved from the server and that compare with the protected URL by optical semblance assessment that is compare the content of the both suspected and protected URL in Block, Layout and over all similarity levels.

In addition to content, style consistency is an important feature that can easily trick phishing victims. Experience shows that users generally ignore the detail (textual or graphical) differences between real pages and phishing pages.This result shows that our similarity assessment metrics are suitably defined and compatible with human visual perception.

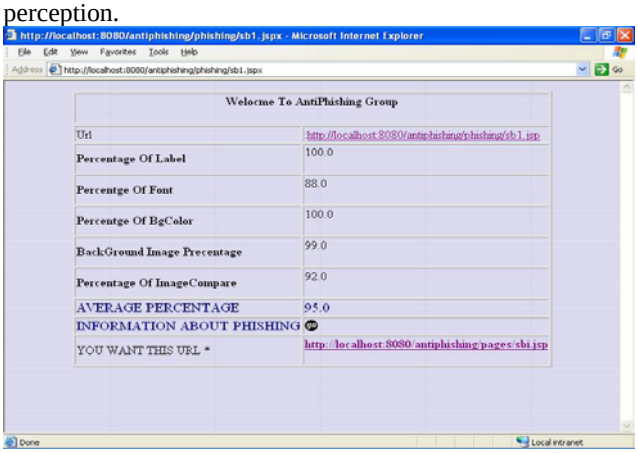

Figure 3.2: Report Identifying the suspected URL.

4. CONCLUSIONS

A novel approach to phishing WebPages strategy based on optical semblance assessment . The flak first fragmentize the webpages into salient blocks according to optical cues. The optical semblance 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. After successfully detecting Suspicious URLs in emails via keywords, we moved on to testing optical semblance assessment of Web pages. When the similarity between a suspicious page and the query exceeds a threshold value for any of the three metrics, the system reports the page as a probable phishing page. To compare similarity based on pure text features (keyword vectors), we also conducted the experiments calculating similarity based on the pages' TF/IDF (term frequency/inverse document frequency) weighted keyword vectors. Our experimental results indicate that the approach can successfully detect phishing Web pages.

This report presents the underlying framework for our optical approach to antiphishing, a component technology for visual similarity assessment of Web pages. 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.

REFERENCES

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

[2] Liu Y., Liu W., & Jiang C. "User interest detection on webpages for building personalized information agent", In Proc. of the 5th International Conference on Web-Age Information Management (WAIM 2004), Dalian, China. LNCS, Vol. 3129, pp. 280–287, 2004.

[3] Chowdhury A., Frieder O., Grossman D., and McCabe. M. Collection statistics for fast duplicate document detection. ACM Trans. on Information Systems (TOIS) 20(2): 171–191, 2002.

[4] Hoad T.C. and Zobel J. Methods for identifying versioned and plagiarised documents. Journal of the American Society for Information Science 54(3): 203–215, 2003.

[5] Broder A., Glassman S., Man asse M., and Zweig G. Syntactic clustering of the web. Proc. of WWW97, pp.391– 404.

[6] Salton G., Wong A., and Yang C.S. A vector space model for information retrieval. Journal of the American Society for Information Science 18(11), pp. 613–620, 1975.

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

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

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

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