Analysis on Visual similarity based antiphising

Prof. Jahnavi Vithalpura
Assistant Professor, IT department,
LD College of Engineering, Gujarat, India
Jahnavi.vithalpura@ldce.ac.in

Ms Niyati Raj
PG student, IT department,
LD college of Engineering, Gujarat, India
Raj.niyati.811@ldce.ac.in

Abstract—Phishing attack uses fake web pages which pretending to be an originalt website and retrieve user information such as bank account noumber, credit card number, passwords and other sensitive details so Anti Phishing is very important for on financial transactions and user privacy protection.
In this paper, I have studied on different methods of phishing detection based on visual similarity and also compared them to check better accuracy and correctness with law performance head.

Keywords—Phishing detection, Visual similarity, Privacy protection

I. INTRODUCTION

Phishing is a form of attack in which attacker steal the sensitive & confidential information such as credit card numbers and passwords with fake web pages. It should be done through emails that misguide users to visit fraud websites that collects users confidential information . Nowadays because of social media usage booming phishing detection is issue of research.

Phishing pages need to lure users by their visual appearance. Page contents and page layouts are visually similar to the original pages. In a web based phishing attack, attacker sets up phishing web pages to lure users to input their sensitive information. The attacker sends emails or publishes web links on social networks that make users to visit phishing pages. Nowdays social networks become a suitable platform to initial social engineering attacks.

Phishing can be detected by analysis of URLs of pages and by page content similarity. Attackers have flexibility in changing URL features to evade detection. One key feature of phishing pages is that they usually use the similar visual appearance as their target pages. The software classification approaches can automatically detect the phishing messages by using white list/blacklist, URL based and Content based.

The black/white-list method is the most widely deployed anti phishing techniques used in browsers. The black/white list methods utilize a blacklist consisting of previously detected phishing URLs, IP addresses or keywords to classify the web page being visited as legitimate or phishing. White list can also be used to filter the famous legitimate web pages.

The most widespread blacklists are the Google safe browsing API[4] and the PhishTank blacklist[5]. Though the blacklist and whitelists are frequently updated, they can not deal with zero-hour phishing attacks[6] because the new zero-hour phishing site can not be added to the blacklist before it is submitted by a victim. The heuristics based methods explore some heuristics that exist in phishing attacks in reality.

In content based detection scheme based on the visual similarity of content between a page and other target pages. The features used include: text and styles, images in the page, and the overall visual appearance of the page. Content based approaches generally extract content features of web pages to identify suspicious websites. To deal with such evasion attempts, some solutions compare images of rendered pages to evaluate their visual similarity.

II. LITERATURE STUDY: TECHNIQUES OF PHISHING DETECTION BASED ON VISUAL SIMILARITY OF WEB PAGES

A. Phishing Alarm: Robust and efficient phishing detection via page component similarity

In this Paper, they have implemented Google Chrome browser extension as shown in Figure 1, the extension consists of three modules: Pre-Processor, Similarity Checker and Target List. The Pre-Processor contains three components: DOM Extractor CSS Extractor, and Visual Characteristics Filter.

CSS Extractor: extracts internal CSS rules directly from the code of web page, and downloads CSS rules in external style sheets from online servers.

DOM Extractor: is in charge of copying the structure of page's body, and acquires the area as well as the value of display and visibility property of each page element.

Visual Characteristics Filter : uses information from both two extractors to exclude CSS rules that have no significant visual influence.

At last , all the rest of CSS rules will be converted into the comparison unit representation and sent to the similarity checker.

Target List stores the comparison units and the URLs of a set of legitimate websites. They include legitimate web pages that

are most likely to be attacked by phishing attackers into the target list. Similarity Checker includes a Similarity Calculator, which computes the similarity value between two pages, and a Decision Maker, which decides whether the suspicious page is phishing or not.

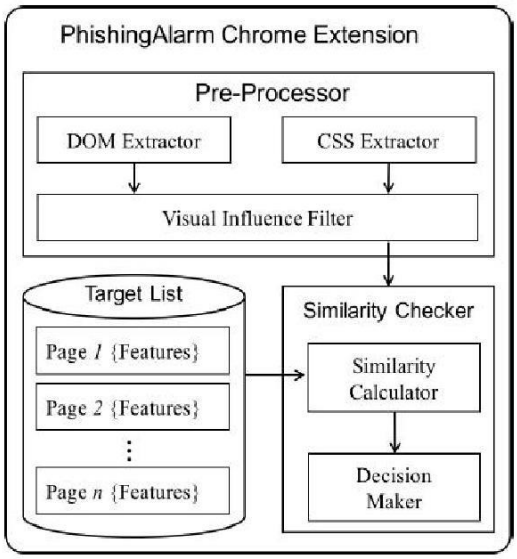

Figure 1. Architecture of Phishing Alarm [1]

Similarity Calculator computes the similarity value between suspicious page and legitimate pages from Target List, one pair at a time. The results are sent to Decision Maker. If any of similarity value is beyond the preset threshold, the suspicious page is classified as phishing.

Experiment Result:

Jian Mao1[1] collected 9,307 verified phishing websites from *PhishTank.com* as the experiment sample set, which consists of phishing pages targeting Apple,PayPal, eBay, and other popular commercial website. They used 6,192 pages from them as the sample set for similarity threshold training, and 3,115 of them as the sample set for correctness assessment (training). To make sure whether these pages are appropriate for their experiment, they manually checked these entire collected sample combined with page element extraction operation. There are 4,934 pages within the training set that cannot be accessed or showed as blank pages, and 2,826 pages of evaluation set had the same problem as well. So they excluded these invalid web pages from their experiment. Besides, they used 46 legitimate pages to test the false positive rate of Phishing-Alarm.

B. Visual Similarity based Anti-phishing with the combination of Local and Global Features[2]

Yu Zhou proposed a novel visual similarity based phishing detection method purely on image level by combining global and local features of the Web page image pair.

The global image feature is extracted only in the visible region of the whole Web page, not in the overall Web page.

The flowchart of their proposed methodology is illustrated in Fig.2, which includes two steps.

The first step is logo detection. First, the snapshot of the suspected Web page and the logo image of the protected Web page are input.

In each image, the Speeded Up Robust Features (SURF) [7] detector is used to detect key points which represent the characteristics of the corresponding image.

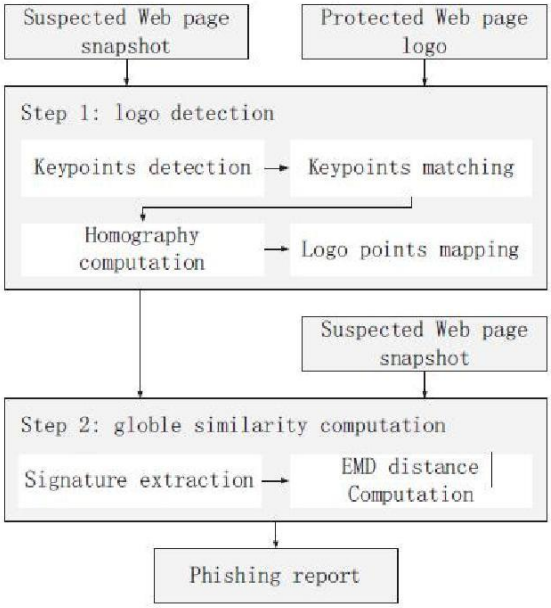

Figure 2. Flow chart of proposed approach [2]

Then, the SURF descriptor is generated for each image. These two sets of key points are matched according to the Euclidean distance. The matched key point pairs are then filtered, and good matched points are reserved. Based on the good matched key points, if the suspected Web page contains the target logo, a homography matrix can be found and the region that the logo locates can be extracted.

The second step is the global similarity computation. The suspected Web page snapshot and the protected Web page snap shot are cut to the visible regions, and two images correspond to the visible regions are obtained. For each result image, they follow the work to extract signature, and the EMD distance between two signatures is taken as the global similarity score. If the suspected Web page snapshot contains the logo of the protected Web page and the global similarity score is beyond to the threshold, the suspected Web page is classified as the phishing Web page. So by this way the local and the global similarities are combined sequentially. In the next two Sections, the logo detection and the global similarity computation are respectively introduced in detail.

Experiment Result :

The experimental results are summarized that the TP rates are all over 90.00%, and the TN rates are all over 97.00%, which proves the effectiveness of the proposed approach.

C. BAIT ALARM : DETECTING PHISHING SITES USING SIMILARITY IN FUNDAMENTAL VISUAL FEATURES [3]

In this paper, Jian Mao, Pei Li proposed a solution, Bait Alarm, to efficiently detect phishing web pages. Page layouts and contents are fundamental feature of web pages' appearance. Since the standard way to specify page layouts is through the style sheet (CSS), they developed an algorithm to detect similarities in key elements related to CSS.

They implemented Bait Alarm in a Google Chrome extension. The overall architecture of the Bait Alarm extension is shown in Figure 3.

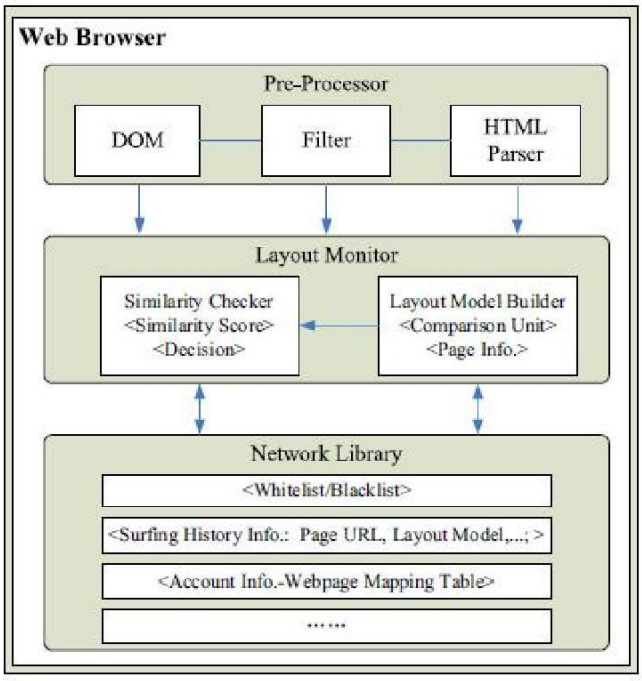

Figure 3. Architecture of Bait Alarm [3]

Bait Alarm includes three main components: Pre-Processor, Layout Monitor and Network Library.

The Pre-Processor consists of *Page Filter*, *DOM*, and *HTML Parser*. After a web page is loaded, the Page Filter checks it over. If the web page has been loaded before, it does not need further analysis. If the loaded page is new and contains some specific UI (e.g., login form), the Page Filter triggers the detecting process. The HTML Parser and the DOM extract the layout information of the suspicious page. When the user inputs personal information, such as Login ID, the browser holds the page and the Pre-Processor sends the layout information to the Layout Monitor. The Layout Monitor consists of a *Layout Model Builder* and a *Similarity Checker*. When the Layout Monitor gets the layout information of the suspicious page from the Pre-Processor, the Layout Model Builder models them into *"comparison-unit"* and sent them to the Similarity Checker, together with additional page features (e.g., page domain, etc.). After the Similarity Checker gets the *comparison unit* of the suspicious page, it searches the Network Library for the victim pages feature model (comparison unit) indexed by the same personal information that has been inputted by the user before.

If the Similarity Checker does not find the matched page, then it informs the browser to release the page and treat it as a new registering web site. The Similarity Checker reports the page information and its layout model to the Network Library.

If the Similarity Checker finds the matched page and gets layout model and additional page information. The checker calculates the similarity score of the pages and outputs the decision based on their similarity score and additional page information. In this scheme, if a page's similarity score is less than the preset threshold, the page is innocent. Then browser releases the page and the Similarity Checker reports the page information and its layout model to the Network Library. Otherwise, the Similarity Checker checks additional page information to make the decision. The checker will submit the related information to the Network Library and inform the browser to pop up a warning page. The Network Library maintains the user's surfing history information (e.g., URL, layout model, etc.), Whitelist/Blacklist and a ―Personal Info-Historical Page Mapping Table‖. The table is used to search for the victim pages based on users' information captured by the browser.

Experiment Result :

To evaluate the accuracy of Bait Alarm, they selected 300 phishing pages from phishtank.com that are labeled as phishing pages of Google, Hotmail, ASB Bank and Blizard corporation respectively. Bait Alarm filtered 149 phishing pages that were not visually similar to the target page claimed by phishtank.com. The other phishing pages remained are checked out by BaitAlarm successfully. For all these 300 testing samples, the detection rate of BaitAlarm is 100% and false negative rate is 0%.

III. COMPARATIVE STUDY

Methods	Phishing Alarm	Visual Similarity	Bait Alarm
Compare CSS	√	-	√
Compare Web page Image	-	√	-
Compare Logo	-	√	-
Merits	More efficient Than Visual Similarity approach and Bait Alarm	Compare Logo and Visual part of web page Only this method can deal with the no text tricks with logo	Works on CSS Give faster result than image matching process

IV. CONCLUSION

Phishing attack popularly used by attackers to collect sensitive information from users. We have reviewed paper on phishing attack based on visual similarities and from that we found that CSS based phishing detection are more efficient and faster as compared to other methods.

V. REFERENCES

[1] Jian Mao1, (Member, Ieee), Wenqian Tian1, Pei Li1, Tao Wei2, (Member,Ieee), And Zhenkai Liang3, (Member, IEEE)"Phishing-Alarm: Robust and Efficient Phishing Detection via Page, 2017. Component Similarity" in special section on privacy preservation for large-scaleUser data in social networks, IEEE Access

[2] Yu Zhou, Yongzheng Zhang , Jun Xiao, Yipeng Wang, Weiyao "Visual Similarity based Anti-Phishing with the Combination of Local and Global Features in" IEEE 13th Conference on Trust, Security and Privacy in Computing and Communications, 2014.

[3] Jian Mao, Pei Li , Kun Li, Tao Wei, and Zhenkai Liang BaitAlarm: "Detecting Phishing Sites Using Similarity in Fndamental Visual Features in" fifth conference on Intelligent Networking and Collaborative Sysytems, 2013.

[4] Google, https://developers.google.com/safe-browsing.

[5] PhishTank, https://www.phishtank.com.

[6] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang, "An empirical analysis of phishing blacklists," in *Sixth Conference on Email and Anti-Spam*, 2009.

[7] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, Speeded-up robust features (SURF), *CVIU*, vol. 110, no. 3, pp. 346 359, 2008.