© 2018 IJSRSET | Volume 4 | Issue 2 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
National Conference on Advanced Research Trends in Information and
ıg Technologies(NCARTICT-2018), Department of IT,L. D. College of Engineering,
Ahmedabad, Gujarat, India In association with
International Journal of Scientific Research in Science, Engineering and Technology

A Survey on Phishing Detection based on Visual Similarity of web pages

Ms Niyati Raj[1], Prof. Jahnavi Vithalpura[2]
[1]PG Student, IT Department, L.D. college of Engineering, Ahmedabad, Gujarat, India
[2]Assistant Professor, IT Department, L.D. College of Engineering, Ahmedabad, Gujarat, India

ABSTRACT

Phishing attack uses scam web pages which pretending to be an important website and takes user's personal information such as credit card number, passwords and other sensitive details. Anti Phishing is very important for online transactions and user privacy protection.
In this paper, I have done survey on different methods of phishing detection based on visual similarity and also compared them to see better accuracy and correctness with law performance head.
Keywords: Phishing detection, Visual similarity, Privacy protection

I. INTRODUCTION

Phishing is a form of social engineering attack in which attacker steals the sensitive information such as credit card numbers and passwords with spoofed web pages. Such communications are usually done through emails that trick users to visit fraudulent websites that collects users' private information. Users use social networks to communicate and share. User's privacy protection has become one of the most research issues.

Phishing pages need to lure users by their visual appearance. Page contents and page layouts are visually similar to the target 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 trick users to visit phishing pages. As social networks become a convenient platform to initial social engineering attacks.

Phishing can be detected by analysis of URLs of phishing pages and by analysis of page content similarity. Attackers have flexibility in changing URL features to evade detection. One key feature of phishing pages is that they usually maintain 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 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 REVIEW

A. Visual Similarity based Anti-phishing with the combination of Local and Global Features[1]

The algorithm is proposed by the Yu Zhou, Yongzheng Zhang , Jun Xiao, Yipeng Wang, Weiyao Lin, year 2014.

In this paper, they 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 approach is illustrated in Figure 1, 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. 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.

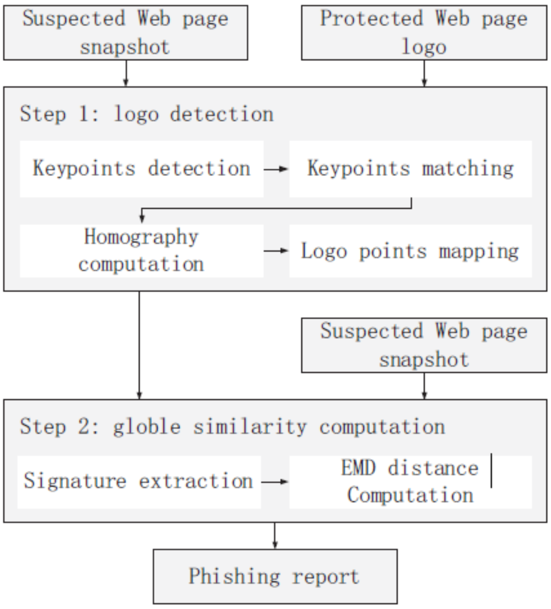

Figure 1. Flow chart of proposed approach [1]

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. In other words, 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.

B. Bait Alarm: Detecting Phishing sites using similarity in fundamental visual features[2]

This algorithm is proposed by Jian Mao, Pei Li, Kun Li, Tao Wei, and Zhenkai Liang, year 2013.

In this paper, they 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.

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.

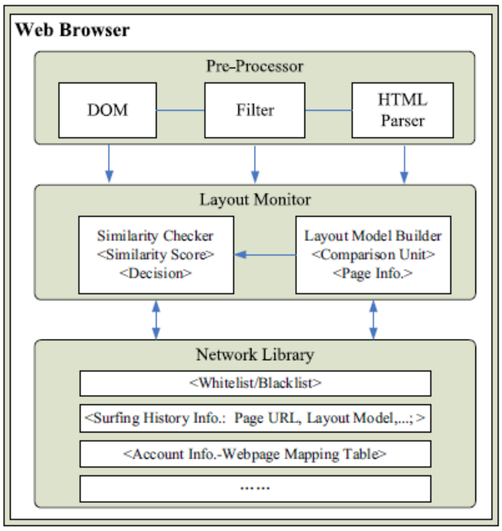

Figure 2. Architecture of Bait Alarm

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.

C. Utilisation of website logo for phishing detection[3]

This algorithm is proposed by Kang Leng Chiew, Ee Hung Chang, San Nah Sze,Wei King Tiong, year 2015. Even though anti-phishing methods based on textual elements receive more attention, it has some limitations. Using a graphical element, especially the logo, is important. This will compensate for the limitations faced in textual-based methods, and will make the detection more robust. They use a logo image to determine the identity consistency between the real and the portrayed identity of a website. Consistent identity indicates a legitimate website and inconsistent identity indicates a phishing website.

The proposed method consists of two processes, namely logo extraction and identity verification. The first process will detect and extract the logo image from all the downloaded image resources of a webpage. In order to detect the right logo image, they utilise a machine learning technique. Based on the extracted logo image, the second process will employ the Google image search to retrieve the portrayed identity.

Since the relationship between the logo and domain name is exclusive, it is reasonable to treat the domain name as the identity. Hence, a comparison between the domain names returned by Google with the one from the query website will enable them to differentiate a phishing from a legitimate website.

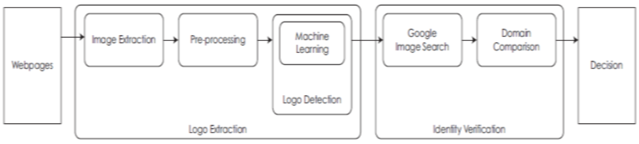

Figure 3. Framework of System [2]

The proposed method involves two main processes: logo extraction and identity verification. A logo extraction process will extract the logo from the query website. Based on the extracted logo, the identity verification process will evaluate the consistency between the real identity and portrayed identity of the query website. If the identity is consistent, the query website is legitimate, and vice versa.

Consistent identity means that the real identity and the portrayed identity are identical. The real identity can be obtained from the domain name of the query website. Whereas the portrayed identity can be retrieved from a database which entry has logo matches to the extracted logo.

Since the relationship between the logo and domain name of a website is exclusive, any mismatch is an indication to the phishing attack. Clearly, a complete and up-to-date database of different website logos with the corresponding domain names is needed. Maintaining this database effectively alone is impossible. Hence, they utilised Google Images as a source of the knowledge database. The first page of the search result. Next, a domain comparison sub process will parse from the first and third elements of the search result. After that, this sub process will extract only the domain name from each of the URLs and compare them to the domain name of the query website. They have taken the liberty to refer to the name which excludes the TLD (top level domain) and any sub domain as the domain name. For example, the domain name for http://www.mydomain.com is my domain. If the comparisons return at least one match, this method will classify the query website as legitimate. Otherwise, it is classified as a phishing website.

As for the limitation in the logo detection sub process mentioned above (i.e., when multiple images of logo and non-logo are returned), this method will repeat the identity verification process for each image, and aggregate the comparison results. Similarly, this method will classify the query website as legitimate if the aggregated comparisons return at least one match.

To fully utilise the Google Images database, authors employed the content-based image retrieval feature from the Google image search facility. It allowed them to retrieve the portrayed identity of a query website from the vast image database. This is depicted as a Google image search sub process. The output from the Google image search sub process is the search result which includes elements.

D. Use of HOG Descriptors in Phishing detection[4]

This method is proposed by ahmet Selmen Bozkir and Ebru akcapinar Sezer in year 2016.

This paper proposed to evaluate and solve this problem by leveraging a pure computer vision based method in the concept of web page layout similarity. Proposed approach employs histogram of oriented gradients descriptor in order to capture cuse of page layout without the need of time consuming intermediate stage of segmentation.

This system was designed to detect zero – day phishing attacks.

For the following reasons, HOG descriptors were preferred in this study:

(i) HOG descriptors are able to capture visual cues of overall page layout

(ii) They are able to provide a certain degree of rotation and translation invariance.

Extracting HOG descriptors require three main steps:

(i) Gradient computation:
Grid of equal sized cells is obtained by dividing the image.

(ii) Orientation Binning
For each pixel, gradient vector is converted to an angle and orientation bins are built according to angle ranges.

(iii) Block normalization
Normalized histograms are concatenated and final descriptor is formed.

This system consists of two modules. The first module so called "wrapper", was designed and implemented in order to find out effective page boundaries and taking a screenshot of web page.

Second module called as “Hogger” was implemented in order to take JPEG file and output a concatenated HOG feature vector.

a. Identifying Region of Interest:

The “wrapper” window was precisely set for taking 1024 pixel wide screen shots. At next stage, crop the portion below 1024 pixels. For the cases where height of web page is lower than 1024 pixels, apply a dominant color detection method for filling the empty lowest part in order to have full square input image. In this way, input mages were generated concerning the existing dominant color in web page. Finally the output image will be converted to grayscale in order to increase the gradient computation accuracy.

b. Revealing the Cues of Page Layout via HOG Descriptors

In order to reveal the appropriate cell size this system applied two different grid configurations. In first configuration (HOG128), the input image consisting of 1024×1024 pixels was divided into 8×8 cells having side length of 128 pixels.

For the second configuration (HOG64), the side length of square sized cells is reduced to 64 pixels which totally results 16×16 grid.

By use of these two types of grid configuration, it aimed to understand and evaluate the levels of details.

c. Use Case Scenario

First collect URLs of legitimate pages LPi which have potential phishing risk and the layout signature of the LPi is stored in legitimate corpus database along with its root domain. Once all the pages which need phishing detection were loaded to the central corpus, a suspicious page SPj can be checked against the legitimate corpus in order to verify whether it has a high similar legitimate target. During the verification process, Histogram Intersection Kernel (HIK) is employed as a similarity metric.

E. A Computer vision technique to detect Phishing Attacks[5]

The algorithm is proposed by Routhu Srinivasa Rao, Syed Taqi Ali, year 2015, India.

In this paper, they proposed a novel solution to defend zero-day phishing attacks. Their proposed approach is a combination of white list and visual similarity based techniques. They used computer vision technique called SURF detector to extract discriminative key point features from both suspicious and targeted websites. Then they are used for computing similarity degree between the legitimate and suspicious pages.

The basic idea of their proposed solution is described below.

1) Maintain a legitimate image database consisting of all popular website screenshots along with their URLs (whitelist).
2) Obtain the accessed URL and do comparing with whitelist of URLs.
3) If comparison is successful URL will be considered as Innocent and no further checking will be required. This removes the extra overhead of comparison of legitimate website display.
4) If the given URI is not found in the white-list then SURF algorithm is applied on the suspicious website screenshot and legitimate image database.
5) Extract the SURF features from both suspicious website screenshot, image database and compare for similarity check.
   a) If similarity score is greater than the threshold, webpage is considered as suspected.
   b) If similarity score is less than threshold, URL is considered as innocent. The domain will be the part of white-list in next update.

III. COMPARISION

Research Paper No.	Merits	Demerits
1.	Purely work on image level Can achieve over 90% TP rate and 97% TN rate.	Some phishing web pages do not contain the official logo which results in the failure of logo detection and then these pages are classified as normal.
2.	Better than white list based techniques.	Needs to be enhanced.
3.	Use logo images to determine the identity consistent between the real identity and the portrayed identity of a website. The captured screenshot is the actual rendered web content, which means there is no other hidden image.	Need enhancement of the logo extraction process with a more effective logo detection algorithm.
4.	An efficient and fast phishing page detection scheme	Can be enhanced by providing image content invariance.
5.	Combination of white list and visual similarity based technique Used SURF detector	Needs to improve the computational cost and accuracy cost.

IV. CONCLUSION

Phishing is a most popular attack used by attackers to collect sensitive information from users. We surveyed paper on phishing attack based on visual similarities and from that we conclude that CSS based phishing detection are more efficient and faster in compare of other methods.

V. REFERENCES

[1] 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.

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

[3] Kang Leng Chiew, Ee Hung Chang, San Nah Sze,Wei King Tiong "Utilisation of website logo for phishing detection" in Elsevier 2015.

[4] ahmet Selmen Bozkir and Ebru akcapinar Sezer -"Use of HOG Descriptors in Phishing detection" in 4th international symposium on digital forensics and security (ISDFS'16), 2016.

[5] Routhu Srinivasa Rao, Syed Taqi Ali "A computer vision technique to detect Phishing Attacks" 5th international conference on communication systems and network technologies, 2015

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

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

[8] 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.

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