Phishing Information Identification using SVM Algorithm

A. P. Deore

Department of Computer Science,
Marathwada Institute of Technology,
Aurangabad, Maharashtra, India

J. S. Kharat

Department of Computer Science,
Marathwada Institute of Technology,
Aurangabad, Maharashtra, India

Abstract- The Phishing is a harmful technique that is used to steal the information from unskilled computer users. The Phishing badly effect on users information which may includes financial details, credential information. Different types of statistical learning based classification methods are available to differentiate the phishing webpage's from original. Feature selection method is the concept, which has been implemented into development of web phishing information detection technique. Different feature of the respective model can be evaluated by Novel framework where we use content-based anti-phishing technique. Feature fusion strategy is one of the key factors of the framework which collect the result of each model and differentiate phishing and original webpage by using under-sampling classification strategy.

Keywords – Phishing, SVM, Stemming, Web page classification, Feature fusion strategy

I. INTRODUCTION

Phishing is the technique to fool a computer user into submitting personal information by creating a similar website that looks like an original site. The phishing become a criminal activity which leads to financial loss and Personal data. Most of the software provider and researchers work on automated anti-phishing tools. The Various methods for phishing web page detection are classified into automated anti-phishing toolbars, User interface anti-phishing platforms, Content based anti-phishing.

In this paper basically we are concentrating on the Content-based anti-phishing. The content based anti phishing technique is one of the powerful which is include the features of web pages, consists of textual content, surface level characteristics and visual content.

The content based anti-phishing technique works on each and every informational part of the webpage and the information of a webpage can be classify in Surface-level characteristics used by toolbars to detect phishing contents. Ex. the Spoof-Guard [2] is a toolbar, makes use of detect the domain age, logos, Uniform Resource Locator address, and hyperlinks to access phishing web pages. An article [3] proposed the concept of semantic link network (SLN). The phishing target of a given webpage is being automatically detected by SLN. The SLN is designed from first finding the associated web pages of the input webpage.

In this paper a hybrid anti-phishing framework approach has been referring to content based Anti phishing methods. The framework basically work on different aspect, like visual and text content from the input web page and Not only Text classifier by the SVM (Support Vector Machine) rules but also an image classifier using the SVM similarity assessment automatically reports a phishing web page, a SVM method to define the threshold, at the last, the results is being combined by data fusion algorithm . The vector space model [4] is a technique to categorize the document in textual partition. Contents are depicted as below:

$$d= (a_1, a_2.....a_N) \qquad (1)$$

Word i with weight a_i in document d. The way of evaluating the weight a_i is based on two observations related to text:

[1] The more times a word present in a document, the more relevant it is to the topic of the page.

[2] The more times a word present throughout a document in the dataset.

The weight of word is calculated by most popular method TFIDF (Term frequency Inverse document frequency) [1].

$$a_i = TF(\omega_i \rightarrow d_i) * IDF(\omega_i) \qquad (2)$$

TFIDF assigns the weight to word ω_i in document d in proportion to the number of occurrences of the word in the document, and in inverse proportion to the number of documents in the collection for which the word occurs at least once. Firstly separate out words from HTML tag and build vocabulary then apply stemming [5] to each word.

While creating phishing web page phisher maintain high similarity with original webpage with respect to text and visual content such as visual layout, aver all design of page, logos, textual contents. Proposed method to calculate the visual similarity of web pages is need to converted HTML web pages into block images and then applies the earth mover's distance [6] method. This method work on the pixel level of page without considering the text level to detect phishing webpage.

The remaining sections of this paper are organized as follows. In the following section, we have introduced the reference paper in Literature Survey and understand an overview of our framework in Proposed System section and at the end conclusion and future enhancement.

II. LITRATURE SURVEY

Existing system uses the different automatic and content based Phishing technique involving authentication, attack tracing and filtering..

The various methods [7] are proposed to classify phishing web page from trusted such as toolbar-based anti-phishing, user interfaced anti-phishing and web page content-based anti-phishing. Anti-phishing toolbars guides the user how to interact with secured website. The poor design of a website is more susceptible to the phisher attack. Internet browser toolbar is one of the tool to prevent webpage from phishing attack. Some anti-phishing toolbars are having built in features such as Spoof- Guard, Net-craft Anti-phishing toolbar, Google toolbar and Internet Explorer 7.0. The author has explained determined the usability of a Anti-phishing toolbar. Research objective is to find out general usability design principles for anti-phishing client side applications. In the paper, effect of weak usability performance of the toolbars is discussed such as Client side function associated with server, toolbar should not be restricted to phishing prevention and at the end, result of usability evolution has been summarized.

M. Wu, R. C. Miller et al. [8] introduced a new anti-phishing solution, the Web Wallet which is an anti-phishing browser side window that allow user to submit credential information such as login name and password instead of original website. But before redirect to requested webpage, the Web Wallet ensure if the requested site is good enough to accept the credential data The function of web wallet is to search not only user's requested webpage but safe path . If requested webpage is not available , web wallet search for alternative secure connection. Web wallet effectively prevent phishing attack, so it is identify a promising approach in anti-phishing technique.The various case study has been included to clarify the use of web wallet against phishing The study also come up with spoofing attack of web wallet. Again we can conclude that web wallet is not trustworthy tool to deal with phishing attack.

M. F. Porter [5] proposed a procedure for suffix stripping. This paper has implemented algorithm to yield the terms with a common stem. The primary goal of suffix stripping algorithm is, to improve Information retrieval environment and this can be done by removal of different suffixes such as -ING, –ED,,-ION,-IONS to leave single stem.

Y. Zhang et al. [9] proposed a content-based approach to detecting phishing web sites instead of automatic toolbar. In this paper, the design, implementation, and evaluation of CANTINA, based on the TF-IDF information retrieval algorithm, content-based approach to detecting phishing web sites has been presented. They have discussed the design and evaluation of several heuristics and various experiments they developed to reduce false positives. The conclusion of the paper to show that CANTINA is good at detecting phishing websites.

CANTINA works as, for incoming a web page, calculate the TF-IDF scores of each term on that web page. Generate a lexical signature from the five terms with highest TF-IDF weights. Supply this lexical signature to a search engine as a input, which in our case is Google. If the domain name of the current web page is similar to the domain name of the N top search results, we consider it to be a legitimate web site. Otherwise, they consider it a phishing site.

W. Liu et al. [10] proposed an approach to detection of phishing webpages based on visual similarity is proposed. A webpage is reported as a phishing suspect if the visual similarity is higher than its corresponding preset threshold. A user can use this approach to search the Web for suspicious webpages which are visually similar to the original webpage. A webpage is detected as a phishing suspect if the visual similarity is higher than its corresponding preset threshold. The previous experiments show that the approach can successfully detect those phishing webpage for online use.

A. Y. Fu et al. [6] proposed a phishing Web page detection method using the EMD-based visual similarity assessment. This approach works at the pixel level of Web pages rather than at the text level, which can detect phishing Web pages only if they are "visually similar" to the protected ones without considering the similarity of the source codes. Experiments also show that our method can achieve satisfying classification precision and phishing recall and the time efficiency of computation is acceptable for online use.

The usage of the method proposed in this paper may not just be limited to the server sides. We are also working on developing a client-side application, SiteWatcher Client, which can be installed by individual users for phishing detections.

III. PROPOSED ALGORITHM

The content representation of proposed system is divided into two categories.

Textual content: "Textual content" in this paper is defined as the terms or words that appear in a given web page, except for the stop words. We first separate the main text content from HTML tags and apply stemming [5] to each word. The function of stemming process is to find out stems rather original world. For ex., 'work', 'works', and 'working' are stemmed into 'work' and considered as the same word.

Visual content: "Visual content" concern to the features with respect to the block regions, layout, overall style including the logos, images and forms. Visual content also can be further specified to the color of the web page background, the font style, the locations of images, the font size and logos, etc. Moreover the visual content is also user-dependent. On the other hand, let's consider the pixel level web page, whereas an image that enables the total representation of the visual content of the web page.

The anti-phishing approach used in this paper shown in Figure 1 contains the following components.

1. A text classifier using the SVM rules to handle the text content extracted from a given web page.
2. An image classifier using the SVM similarity assessment to handle the pixel level content of a given web page that has been transformed into an image.
3. A SVM approach to estimate the threshold used in classifiers through offline training.
4. A data fusion algorithm to combine the results from the image classifier and the text classifier. The algorithm employs the SVM approach as well.

The system includes a training section, which is to estimate the statistics of historical data, and a testing section, which is to examine the incoming testing web pages. The statistics of the web page training set consists of the probabilities that a textual web page belongs to the categories, the matching thresholds of classifiers, and the posterior probability of data fusion. Through the preprocessing, content representations has been done, i.e., visual and textual, are continuously extracted from a given testing web page. The text classifier is used to classify the given web page into the corresponding category based on the textual features.

The image classifier is used to classify the incoming web page into the relevant category based on the visual content. Then the fusion algorithm combines the detection results generated by the two classifiers. The detection results are eventually transmitted to the online users or the web browsers.

In preprocessing, first step is to separate HTML tags from the main contexts of an incoming web page. We construct a word vocabulary To form a histogram vector for each web page. This system extracts all the words from a given protected web page and applies stemming to each word. The SVM word-based extraction delivers more discriminative information than stemming-based extraction. But point out that the SVM word-based extraction will largely increase the vocabulary size. In addition, using stemming will build more robustness of detection, because phishers may manipulate the textual content through the change of tense and active to passive.

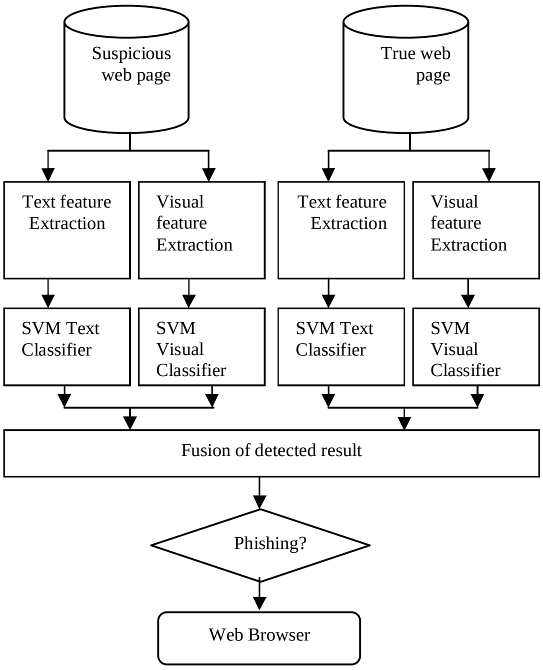

Figure 1. Architecture Design for phishing web page detection system.

While stemming for smaller vocabulary detection and robust detection size, to identify similar textual content, we suggest word-based extraction using the SVM. Given a web page, where each component represents the term frequency and n denotes the total number of components in a histogram vector. We explain three points here.

[1] We do not extract words from all the web pages in a dataset to construct the vocabulary because phishers use the text from a targeted web page to scam users.

[2] For simplicity, we do not use any feature extraction algorithms in the process of vocabulary construction.

[3] We do not take the similar web pages into account because the sizes of most phishing web pages are small.

In reality, using only text content is insufficient to detect phishing web pages. This technique usually leads to high FP (false positives), because phishing web pages are mostly similar to the targeted web pages not only in textual content but also in visual content such as layout, logos, and style. In this system, we use the same approach as in using the SVM to measure the visual similarity between an input web page and a secured web page.

Firstly, we retrieve the vulnerable web pages and secure web pages from the web. Second, we generate signatures of input webpage, which are used for the calculation of the SVM between them. Each and every web page images are normalized into fixed-size box images. We use these normalized images to generate the signature of each web page.

The image classifier is implemented by setting a threshold, which is later estimated in the subsequent section. If the visual similarity between a input web page and the secure web page exceeds the threshold, it means the web page is classified as phishing.

IV.CONCLUSION

The various anti-phishing techniques have been invented over a period for effective search of Phishing webpage. After evaluating study of different anti phishing method, we can conclude that the content based anti-phishing is one of the efficient method. In this system, we have designed a strong framework to detect phishing webpage. This system includes text classifier, Image classifier and fusion algorithm. Text Classifier applies the SVM rules and classifies the webpage in phishing or original whereas Image classifier evaluates visual similarity between incoming webpage and secure webpage. The probabilistic model derived from SVM of both Text and Image classifier produces the matching threshold between input webpage and secure webpage. The role of Data Fusion model is to combine the result of both classifier and display result. This implemented model can easily embed into industrial anti-phishing system. In future work, adding more content features and building more dataset of updated webpage in current model can lead to more accuracy.

REFERENCES

[1] *Global Phishing Survey: Domain Name Use and Trends in 1H2009.*Anti-Phishing Working Group, Cambridge, MA [Online]. Available:http://www.antiphishing.org.

[2] N. Chou, R. Ledesma, Y. Teraguchi, and D. Boneh, "Client-side defense against web-based identity theft," in *Proc. 11th Annu.Netw.Distribut.Syst. Secur. Symp.*, San Diego, CA, Feb. 2005, pp. 119–128.

[3] W. Liu, N. Fang, X. Quan, B. Qiu, and G. Liu, "Discovering phishing target based on semantic link network," *Future Generat. Comput.Syst.*,vol. 26, no. 3, pp. 381–388, Mar. 2010

[4] Baeza-Yates, R. A. and Ribeiro-Neto, B. Modern Information Retrieval.Addison-Wesley Longman Publishing Co., Inc. 1999.

[5] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3,pp. 130–137, 1980.

[6] A. Y. Fu, W. Liu, and X. Deng, "Detecting phishing web pages with visual similarity assessment based on earth mover's distance (EMD)," *IEEE Trans. Depend. Secure Computer*., vol. 3, no. 4, pp. 301–311, Oct.–Dec. 2006.

[7] L. Li and M. Helenius, "Usability evaluation of anti-phishing toolbars," J. Comput. Virol., vol. 3, no. 2, pp. 163–184, 2007.

[8] M. Wu, R. C. Miller, and G. Little, "Web wallet: Preventing phishing attacks by revealing user intentions," in Proc. 2nd Symp. Usable Privacy Secur., Pittsburgh, PA, Jul. 2006, pp. 102–113.

[9] Y. Zhang, J. Hong, and L. Cranor, "CANTINA: A content-based approach to detecting phishing web sites," in Proc. 16th Int. Conf. World Wide Web, Banff, AB, Canada, May 2007, pp. 639–648

[10] W. Liu, G. Huang, X. Liu, M. Zhang, and X. Deng, "Detection of phishing web pages based on visual similarity," in Proc. 14th Int. Conf. World Wide Web, Chiba, Japan, May 2005, pp. 1060–1061

[11] 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.

[12] Liu Y., Liu W., and Jiang C. User interest detection onwebpages 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.