Detecting Phishing Web Pages using NB classifier and EMD Approach

Pranali P. Akare[1]
Student of BE Computer Engineering
BVCOE & RI, Nasik, India
University of Pune
pranali.akare4@gmail.com

Heena.Mohm.H.Maniyar[2]
Student of BE Computer Engineering
BVCOE & RI, Nasik, India
University of Pune
heena.shaikh735@gmail.com

Tejendra D. Thorat[3]
Student of BE Computer Engineering
BVCOE & RI, Nasik, India
University of Pune
thorattejendra1011@gmail.com

Jagruti k. Pagar[4]
Student of BE Computer Engineering
BVCOE & RI, Nasik, India
University of Pune
pagarjagruti@yahoo.com

Abstract— Phishing is online security attack where attacker create replica of exiting web page for accessing users password, personal, financial data. Phishing is a form of online fraudulent activity in which an attacker aims to steal a victim's personal information, such as a bank account, online banking password or a credit card number. Targets are tricked into providing such information by a combination of spoofing techniques and social engineering. We have proposed a new approach named as "Anti phishing framework using Bayesian approach for content based phishing web page detection. Our model used to detect the similarity between suspicious web page and secure web page through image and text contained by the web page. Text classifier, an image classifier and fusion algorithm the result from classifier are introduced. In the text classifier naive Bayes algorithm is used to calculate the Probability, an image classifier the earth mover's distance algorithm is used to measure the visual Similarity and our Bayesian model is designed to determine the threshold. In data fusion detection for image classifier and text classifier means how many web site image and text are matched exactly. If any web page contains above 50% spam text and image so we are declare these web sites are phishing other-wise not phishing.

Keywords— *Bayes theory, classifier, data fusion, phishing detection, web page.*

*****

I. INTRODUCTION

MALICIOUS people, known as phishers create phishing web pages, i.e., forgeries of real web pages, to steal individuals' personal information such as user's password, online bank account number, credit card number, and other financial data unwary online users can be easily deceived by these phishing web pages because of their high similarities to the real ones. The Anti-Phishing Working Group reported that there were at least 55 698 phishing attacks. The latest statistics show that phishing remains a major criminal activity involving great losses of money and personal data. Automatically detecting phishing web pages has attracted much attention from security and software providers used to academic researchers. Methods for detecting phishing web pages can be classified into industrial toolbar based on anti-phishing, web page content-based anti-phishing and user-interface-based anti-phishing[3].

To date, techniques for phishing detection used by the industry mainly include filtering, attack analyzing and tracking, authentication, phishing report generating, and network law enforcement. These anti-phishing internet services are built into e-mail servers and web browsers and available as web browser toolbars (e.g., Spoof Guard Toolbar1, Trust Watch Toolbar2, and Net craft Anti-Phishing Toolbar3). These industrial services, however, do not efficiently thwart all phishing attacks[6]. Conducted thorough study and analysis on the effectiveness of anti-phishing toolbars, which consist of three security toolbars and other mostly used browser security indicators. The study specifies that all examined toolbars in were ineffective to prevent web pages from phishing attacks.

Reports show that 20 out of 30 subjects were spoofed by at least one phishing attack, 85% of the spoofed subjects specified that the websites look legitimate or exactly same as they visited before, and 40% of the spoofed subjects were tricked due to poorly designed web sites. Cranor et al. [7] performed another study on an evaluation of 10 anti-phishing tools. They indicated that only one tool could consistently detect more than 60% of phishing web sites without a high rate of false positives, whilst four tools were unable to recognize 50% of the tested web sites[1]. Apart from these studies on the effectiveness of anti-phishing toolbars, investigated usability of five typical anti-phishing toolbars. They found that the main user interface of the toolbar, warnings, and help system are the three basic components that should be well designed. They also found that it is beneficial to apply whitelist and blacklist methods together.

They also found that it is beneficial to apply white list and blacklist methods together. Also, due to the quality of the online traffic the applications from the anti-phishing client side should not rely merely on the Internet[5]. Aburrous recently developed a resilient model by using fuzzy logic to quantify and qualify the website phishing characteristics with a layered structure and to study the influence of the phishing characteristics at different layers on the final phishing website rate.

II. LITERATUR SERVEY

Phishing is an attempt by an individual or a group to thieve personal confidential information such as passwords, bank account number, credit card information etc from unsuspecting victims for identity theft, financial gain and other fake activities. In this paper we have proposed a new approach named as”A Novel Ant phishing framework based on visual cryptography” to solve the problem of phishing web page. An image based verification using Visual Cryptography (vc) is used here[5]. The visual cryptography is used to explored preserve the privacy of image captcha by decayed the original image captcha into two shares that are stored in separate database servers like the original image captcha can be revealed only when both are concurrently available; the individual sheet images do not reveal the identity of the original image captcha. Once the original image captcha is revealed to the user it can be used as the password. Online transactions are nowadays become very common and there are various attacks present behind this. In these types of various phishing, attacks is identified as a major security risk and new innovative ideas are arising with this in each second so precautionary mechanisms should also be so effective [4].

These types of standard technologies have several disadvantages:

1. Blacklist-based technique with low false alarm probability, but it cannot distinguish the websites that are not in the blacklist database[3]. Because the growth of phishing websites is too short and the establishment of blacklist has a long delay of time, the accuracy of prohibit is not too high.

2. Heuristic-based anti-phishing technique, with a high probability of false and fails alarm, and it is easy for the defender to use technical means to avoid the heuristic characteristics detection.

3. Similarity assessment based technique is time-consuming. It needs too long period to calculate a pair of pages, so use this technique to notice the phishing websites on the client terminal is not suitable. And there is low rate of accuracy this method depends on many factors such as the images, text and measurement similarity technique. However, this technique (in particular, image resemblance proof of identity technique) is not perfect enough yet[7].

In this paper we are using a Bayesian approach algorithms for content-based web page detection is presented. Our model takes into phishing account textual and visual contents to measure the similarity between the protected web page and suspicious web pages. Also we are use second fusion algorithms for image classifier and text classifiers means how many web site image and text are match exactly if any web page in above 90% match text and image so we are declare these web sites are published other-wise not published[1].

III. SYSTEM OVERVIEW

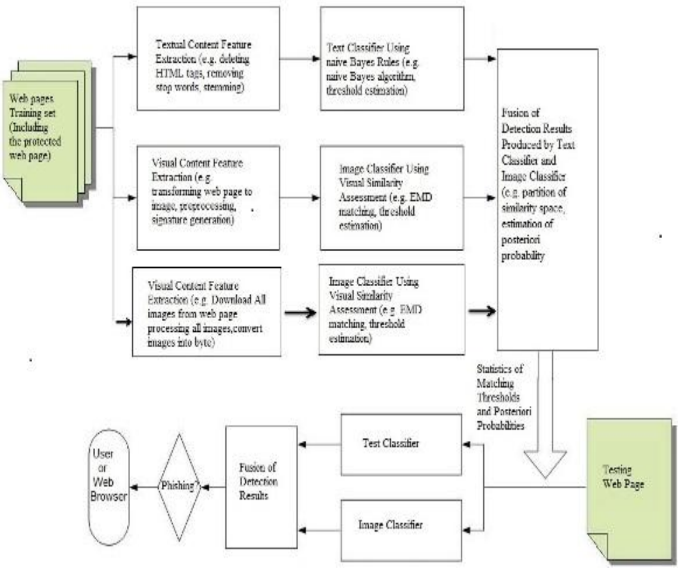
Web pages Training set (Including the protected web page)
Textual Content Feature Extraction (e.g. deleting HTML tags, removing stop words, stemming)
Text Classifier Using naive Bayes Rules (e.g. naive Bayes algorithm, threshold estimation)
Visual Content Feature Extraction (e.g. transforming web page to image, preprocessing, signature generation)
Image Classifier Using Visual Similarity Assessment (e.g. EMD matching, threshold estimation)
Visual Content Feature Extraction (e.g. Download All images from web page processing all images,convert images into byte)
Image Classifier Using Visual Similarity Assessment (e.g. EMD matching, threshold estimation)
Fusion of Detection Results Produced by Text Classifier and Image Classifier (e.g. partition of similarity space, estimation of posteriori probability
Statistics of Matching Thresholds and Posteriori Probabilities
Testing Web Page
Test Classifier
Image Classifier
Fusion of Detection Results
Phishing?
User or Web Browser

Figure 1.System Architecture

As shown in figure 1 system contain different algorithm for detecting phishing.

Textual Content-Based Anti-Phishing:-

Text Classifier Using Nevi Bayesian Approach:-

1. Surface level content. “Surface level content” here is defined as the characteristics that are used by the users to access to a web page or to connect to other web pages.
2. Such surface-level content consists of the domain name, URL, and hyperlinks which are involved in a given web page.
3. 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 (a set of common words like “a,” “the,” “this,” etc.).
4. We first separate the main text content from HTML tags and apply stemming to each word. Stems are used as basic features instead of original words. For example, “program,” “programs,” and “programming” are stemmed into “program” and considered as the same word.
5. Find the probability using Nevi Bayesian Rule.

Visual Content-Based Anti-Phishing:-

Image Processing EMD Approach:-

1. Convert Web page into image. the Microsoft IE browser is used to transform HTML and accessory files on the screen into web page images (in JPEG format).
2. Process on image resize image into specific size (e.g. 200×200).
3. Convert resized image into Gray-Scal.
4. Get Bytes of image.
5. Calculate the EMD and visual similarity between the input web page image and the protected web page image.

IV. ALGORITHMIC STRATEGY

1. Implementation of text classification

a) Posterior probability:

$$P\ (gj|v1,\ v2,\ .\ .\ .\ ,vn) = \frac{P(v1,v2,\dots,vn\ |gj\ )}{P(gj\ )\ (v1,v2,\dots,vn\ )}$$

Where, The prior probability P(gj) is estimated by the frequency of the training samples belonging to category gj.
G = (g1, g2, . . . , gj, . . . , gd) denote the set of web page categories
d = Total number of categories.

b) Conditional Probability:

$$P\ (v1,\ v2,\ .\ .\ .\ ,\ vn\ |gj) = \prod_{i=1}{n}\ \ P(vi\ |gj\ ).$$

Where,P (v1, v2, . . . , vnj gj) denotes Conditional Probability.

c) Calculation of Centroid

$$C\sigma = \sum_{i=1} N\sigma\ (c\sigma, i\ /N\sigma\ )$$

Where, *Cσ*, i = Coordinate of the ith pixel that has the degraded color*σ*.
N*σ* = Total number of pixels that have the degraded color _ (i.e., the frequency).

2. Implementation of Image Classifier:

a) Signature of an image

S = (F*σ*1, N*σ* 1), (F*σ* 2, N*σ* 2). . . (F*σ* N, N*σ* N)
Where ,N = Total number of selected degraded colors.
F*σ* = *σ*, Cσ be the feature, where *σ* represents the degraded color (i.e., a 4-tuple [A, R, G, B] in which the components represent alpha, red, green, and blue, respectively).

The image classier is implemented by setting a threshold θV, which is later estimated in the subsequent section. If the visual similarity Visual between a suspected web page and the protected web page exceeds the threshold θV , the web page is classified as phishing, Otherwise, the web page is classified as normal. The overall implementation process of image classifier is summarized as follows.

Step 1: Obtain the images of web pages from its URL and perform normalization.
Step 2: Generate visual signature of the input image including the color and coordinate features.
Step 3: Calculate the EMD and visual similarity between the input web page image and the protected web page image.
Step 4: Classify the input web page into corresponding category according to the comparison of the visual similarity and the threshold θ V.

3. Implementation procedures of fusion algorithm

a) EMD

$$EMD\ (Sa,\ Sb,\ D) = \frac{\sum_{i=1} m \sum_{j=1} n f i\ j\ \cdot ji\ j}{\sum_{i=1} m \sum_{j=1} n f i\ j}$$

Where, a and b are two web page images with signature Sa and Sb, respectively,
D = [dij], Distance matrix (1< i < m, 1< j < n),
Where dij = Dnorm (Fσi, F*σ* j).

b) Hybrid Similarity Measurement

S visual (Sa, Sb) = 1 − *(E M D (Sa, Sb, D)) α*

Where _ _ [0,1] is a weighting parameter that is used to balance the weights of similarity measurements from text and image classifier.

Si,T denote the probability that the ith web page belongs to the phishing category associated with the text classifier.
Si, v denotes the similarity of the ith web page and the protected web page.
Posterior probability conditioning on a sub-interval lt = [Lt-1, Lt] for a classifier

$$PT\ (C|lt) = \frac{PT\ (C)PT\ (lt\ |C)}{PT\ (C)PT\ (lt\ |C) + PT\ (I\ )PT\ (lt\ |I\ )}$$

Where,Si,T and Si,V are in the range of [0, 1]
Si,T ∈ [*Lt*−1,*Lt*] (*t* = 1,2,. . ., *L*)
lt = [Lt-1, Lt] = Sub-interval

Step 1: Input the training set, train a text classifier and an image classifier, and then collect similarity measurements from different classifiers.
Step 2: Partition the interval of similarity measurements into sub-intervals.
Step 3: Estimate the posterior probabilities conditioning on all the sub-intervals for the text classifier.
Step 4: Estimate the posterior probabilities conditioning on all the sub-intervals for the image classifier.
Step 5: For a new testing web page, classify it into corresponding category by using the Text classifier and the image classifier. If it is classified into different categories, locate the sub-interval that the similarity measurement of the web page belongs to and execute
step6), if else, execute
Step 7: Calculate the decision factor for the testing web page.
Step 8: Return the final classification results to a user or a web browser.

Table 1 contains description about algorithms as follows:

Table 1: Algorithmic specification

Symbol	Meaning
P	Probability
N*σ*	Number of pixels
Sa,Sb	a,b are two web pages
A	Set of attributes possesses
dij	Distance materix
Si,T	Probability of ith web page
G	Set of web page categry attribute
d	Number of categries
Cσ	Coordinate of the pixel
F*σ*	Features where *σ represents*
N	Numbers of selected degraded colors
lt = [Lt-1, Lt]	Sun interval

V. CONCLUSION

A new content-based anti-phishing system has been developed. In this paper we develop a new framework to solve anti-phishing problem. The new features of this framework is developed by using image classifier, text classifier and fusion algorithm. Based on given textual content, text classifier is able to classify whether given web page is phishing or normal. This text classifier is developed by using naive Bayes rule. Based on given visual content, the image classifier, which is build up by using EMD, is able to differentiate between phishing web page and normal web page. The matching threshold is used in both text classifier and image classifier. Fusion algorithm is used by using Bayesian theory. Our results corroborated the effectiveness of our proposed system. Experimental results suggested that our proposed system is able to improve accuracy of phishing detection. Although the promising results presented in this paper, our future work will to include more features into current model.

REFERENCES

[1] A. Emigh, \xcvxcvxvxc," in Radix Laboratories Inc., Eau Claire, WI [Online]., 2005.

[2] L. G. A. L. P. Brockett, R. Derrig and M. Alpert, \Fraud classi_cation using principal component analysis of ridits," in 7th IEEE International Conference on Computer and Information Technology, 2002, pp. 341{371.

[3] R.Caruana and A. Niculescu-Mizi, \Data mining in metric space: An empirical analysis of supervised learning performance criteria," in 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2004.

[4] P. Christen and K. Goiser, Quality and Complexity Measures for Data Linkage and Deduplication. Quality Measures in Data Mining, F. Guillet and H. Hamilton, 2007, vol. 43.

[5] D. P. C. Cortes and C., \Computational methods for dynamic graphs," J. Computational and Graphical Statistics, vol. 12, 2003.

[6] R. C. A. Goldenberg, G. Shmueli and S. Fienberg, \Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales," Proc. Nat'l Academy of Sciences USA, vol. 99, pp. 5237 {5240,2002}

[7] K. C. G. Gordon, D. Rebovich and J. Gordon, Identity Fraud Trends and Patterns: Building a Data-Based Foundation for Proactive Enforcement. Center for Identity Management and Information Protection, Utica College, 2007

Pranali P. Akare she is BE student of Information Technology Engineering at Brahma Valley College of Engineering And Research Institute, Nasik under University of Pune. She is interested in java programming.

Heena M.H Maniyar she is BE student of Information Technology Engineering at Brahma Valley College of Engineering And Research Institute, Nasik under University of Pune. She is interested in oracle.

Jagruti k. Pagar she is student of Information Technology Engineering student at Brahma Valley College of Engineering And Research Institute, Nasik under University of Pune. She is interested in .net.

Tejendra D. Thorat he is student of Information Technology Engineering student at Brahma Valley College of Engineering And Research Institute, Nasik under University of Pune. he is interested in oracle.