Logo Image Based Approach for Phishing Detection

Himani Thakur[1], Supreet Kaur[2]
[1]M.Tech,Computer Science Engineering Department,
Punjabi University Regional Centre for Information Technology and Management,Mohali
thakurhimani3@gmail.com
[2]Assistant Professor, Computer Science Engineering Department,Punjabi University Regional Centre for Information Technology and Management,Mohali

skaur.gujral@gamil.com

ABSTRACT

Phishing is a cyber attack which involves a fake website mimicking the some real legitimate website. The website makes the user believe the website being authentic and thus online user provides their sensitive information like password, PIN, Social Security Number, and Credit Card Information etc. Due to involvement of such high sensitivity information, these websites are a huge threat to online users and detection and blocking of such website become crucial. In this thesis, we propose a new phishing detection method to protect the internet users from such attacks. In particular, given a website, our proposed method will be able to detect between a phishing website and a legitimate website just by the screenshot of the logo image of it. Due to the usage of screenshot for extracting the logo, any hidden logo will not be able to spoof the algorithm into considering the website as phishing as happened in existing methods. In first study focus was on dataset gathering and then the logo image is extracted. This logo image is uploaded to Google image search engine using automated script which returns the URLs associated with that image. Since the relationship between logo and domain name is exclusive it is reasonable to treat the logo image as identity of original URL. Hence the phishing website will not have the same relation to the logo image as such and will not get returned as URL by Google when search for that logo image. Further, Alexa page rank is also used to strengthen the detection accuracy.

Keywords: Anti-phishing,Website logo,Google image search.

INTRODUCTION

Phishing is the act of mimicking a trusted website to gain sensitive information from online users like detail of credit card, personal identification number etc. Since APWG reports claim that 40-50% of phishing attacks are based on common legal web sites, we decided to check this and so we compiled a list of target words which included many popular phishing targets, such as Ebay and paypal [20]. In most of cases criminals make web pages by copying legitimate or make a little change in page content to gain user’s sensitive information.For example, a system can be technically secure enough against password stealing, however uninformed end users if click on the Hypertext Transfer Protocol (HTTP) link may leak their passwords, which ultimately threatens the overall security of the system. There are many solution exist to detect phishing attack but no one bullet proof solution yet to present which detect all type of phishing attack. In order to detect whether the website is phishing website, the first question to ask is:how to discriminate phishing website and the legitimate website as the reason is that the phishing website is look alike to the legitimate website.Where if the we have the portrayed identity of the query website then we can find out that if it a legitimate or a phishing website. (if the doubt website is a phishing website, the portrayed identity will be the identity of the besieged legitimate website)[1],we can then differentiate the phishing website from the legitimate website. Knowing that the phishers will use the optical factors ripped off from the legitimate website, especially the logo, in their phishing websites, this inspires to propose an anti-phishing method based on the recognition of website identity through the logo. This is rational, as the logo usually symbolizes the identity of a legitimate website.

Query website: The website which is under test to check if it is a phishing website or a legitimate website.Portrayed identity: The trademark or entity for which a legitimate website is emphasize to.As for example, the domain http://www.ebay.com where the portrayed identity of the legitimate website is ebay. Likewise, for e.g., domain http://www.www1-ebaee.com is a phishing website which mimic the ebay website,where the portrayed identity is ebay.Real identity: That is the actual identity of a query website. For example,the domain http://www.ebay.com where ebay is the real identity of a legitimate website. Whereas the domain http://www.www1-ebaee.com is a phishing website which mimic the ebay website, its real identity is www1-ebaee. There are various type of phishing such as email phishing,malware based phishing,keylogger and screenloggers(perticuler type of malware that track the keyboard input and send the relivant information to hacker via the internet),man in the middle phishing,etc. There have been several anti phishing technique developed in last few years.

Related work

While there exists numerous different techniques in phishing detection.These are as following:

Kang Leng Chiew et.al [1]: Used a logo image to find out the identity reliability between the real and the portrayed identity of a website. reliable identity point towards a legitimate website and incompatible identity point towards a phishing website. The proposed technique consists of two procedures, that is logo extraction and identity verification. The first procedure will identify and take out the logo image from all the downloaded image resources of a webpage. In order to identify the right logo image, the method make use of a machine learning technique. Based on the take out of a logo image, the second procedure will utilize the Google image search to retrieve the portrayed identity. Since the connection between the logo and domain name is special, it is logical to treat the domain name as the identity. Hence, a difference between the domain name returned by Google with the one from the query website will allow us to distinguish a phishing from a legitimate website. The carry out experiments show consistent and promising results. This proves the efficiency and possibility of using a graphical element such as a logo to identify a phishing website.

J. Hong and L. Cranor et al. [7]: The majority of the proposed techniques in the literature, the unmediated heuristic approach. One of the popular methods is CANTINA. This method will calculate the TF-IDF from the content of a webpage, and produce a lexical signature. The technique will utilize the generated lexical signature to do a web search through the Google search engine. The returned outcome will be used to conclude the authority of a website.Even though this technique can carry out logically well in the finding of phishing.

J. Lee et al. [8]: Proposed a more recent research study on the characteristics-based heuristic approach is the one proposed by the main objective of the technique is to detect the identity of the phishing target when a phishing webpage is identified. The technique is based on the design of a self-organised semantic data model, labeled as the Semantic Link Network which is frequently used in organising web resources. while this method is using different detection mechanisms, its basis starts from the textual elements (i.e., the taking out of hyperlinks, keywords and textual contents for the process of link relations,search relations and text relations, respectively).

Cao Y et al. [9]: Proposed one can gather a list of legitimate URLs. This method is recognized as whitelisting, and it is also a kind of list-based approach. An example of a whitelisting technique is the research proposed by the authors developed an automated technique that maintains and stores a whitelist at the client side.

Prakash P et al. [10]: Proposed a more active and flexible list-based approach is called PhishNet This technique uses several URL variant heuristics to procedure the existing blacklisted URLs and make multiple variant URLs. The produce URLs will form a analytical blacklist. The results explain that it can successfully detect new and old phishing websites. while a list-based approach provides ease in design and is easy to put into operation, keeping the list complete and up-to-date needs great attempt, and always go through from incompleteness.

Tout and Hafner et al. [4]: Proposed one of the popular techniques is blacklisting. Many well-liked web browsers are utilize this approach to detect phishing website in this technique, a query website is checked with a list (i.e., a list is recognized as phishing URLs), which is collected and upholded by some association or organisation. If the checking returns a match, then the website will be labeled as phishing.

Sadia Afroz et al. [23]: Proposed one of the popular techniques is PhishZoo. This paper proposes a phishing detection approach—PhishZoo that utilizes profiles of reliable websites outer shells to detect phishing.The advantage is that it can categorize zero-day phishing attacks and embattled hits against minor sites (such as corporate intranets). A key role of this paper is that it comprise a presentation study and a structure for making use of computer vision techniques in a sensible way.

Zhuang et.al [12]: Proposed a technique that is deliberated and applied an intelligent model for detecting phishing websites. In this model, they take out 10 different types of kind such as heading, keyword and connection text information to stand for the website. Various classifiers are then build stand on these dissimilar features. They proposed a ethical ensemble classification algorithm to join the expect results from different phishing detection classifiers. Hierarchical clustering technique has been working for mechanical phishing categorization. Case studies on great and actual daily phishing websites composed from King soft Internet Security Lab demonstrate that their proposed model outperforms other commonly used anti-phishing methods and tools in phishing website finding.

Bian et.al [14]: Proposed a method to assess the effectiveness of three popular online resources in identifing phishing sites-viz,Yahoo! Inlink data and Yahoo! directory service, Google PageRank system. Their results point towards that these online resources can be used to boost the accuracy of phishing site detection when used in combination with existing phishing countermeasures. The proposed loom involves investigate the following three attributes of a goal site (site being check up): (1) the reliability of the target sitepsilas hosting domain, (2) the reliability of in-neighbor sites that link to the hosting domain, and (3) the connection between the aim sitepsilas web category and its hosting domainpsilas web kind. The abovementioned online resources by themselves are insufficient to concentrate on the phishing attack problem. This approach provide convention on how each of those resources may be included with existing phishing detection techniques to offer a more efficient solution.

Ali et.al in [15]: Proposed a approach of confidentiality in Instant Messengers (IM) by means of Association Rule Mining (ARM) method a Data Mining approach included with Speech Recognition system. verbal skills are acknowledged from words with the help of FFT spectrum analysis and LPC coefficients methodologies. Online criminal's at the present time modified voice chatting technique along with text messages collaboratively or either of them in IM's and squashing out personal information direct to intimidation and barrier for privacy. To facilitate centre of attention on privacy preserving this approach residential and try out Anti Phishing Detection system (APD) in IM's to detect unreliable phishing for text and audio collaboratively.

Tan et.al in [16]: Proposed an anti-phishing method to protect users against phishing attacks in the internet. The scope of this approach study focuses mainly on the detection of phishing websites with English content. In order to encourage users on whom the website claims to be, phishers usually place brand names in different parts of the URL. They oppressed this phishing pattern by conveying weights to words take out from the HTML content, based on their co-appearance at path,hostname and file names of URLs. These weights are then supplementary to their equivalent TF-IDF weights. The most likely words are particular and submitted to Yahoo Search to recover the highest frequency domain name amongst the top 30 search results. A WHOIS lookup is carry out to disclose the vendor behind the selected domain name. A phishing website can be easily illustrious if the vendor of query domain name be different from the owner of domain name returned by the search engine.

Fang et.al in [17]: Proposed a approach of an artificial protected system for phishing detection. The system is to sense phishing emails throughout mature detectors and memory detectors. The memory detectors are produced from the training data set, which consecutively contains the phishing emails up to that time seen by the system. The immature detectors are replicate through the system's mutation procedure. To the best of this approach facts that this is the first time such a system is ever projected. They assumed that the system is more adaptive than any other active phishing detection techniques.

Nguyen et.al in [18]: Proposed an efficient approach for identifing phishing websites foundation on the single-layer neural network. Particularly, the proposed technique calculates the value of heuristics impartially. Then, the weights of heuristic are produced by a single-layer neural network. The proposed technique is assessed with a dataset of 11,660 phishing sites and 10,000 legitimate sites.

Jo et.al in [19]: Proposed a approach to consider websites' identity claims. Their phishing detection system copy this human expert behavior. Given a website, their system study the identity that this website assert, and figure the documentary significance between this claimed identity and other description in the website. Their phishing detection system then employ this textual significance as one of the sort for classification.

DeBarr et.al [3]: Proposed a approach as a first step the exercise of Spectral Clustering to analyze messages based on traffic behavior. specifically, Spectral Clustering analyzes the association between URL substrings for web sites originate in the message contents. Cluster membership is then employ to assemble a Random Forest classifier for phishing. Data from the Phishing Email quantity and the Spam killer Email quantity are used to evaluate this approach. Performance assessment metrics include the region Under the receiver operating characteristic Curve (AUC), as well as accurateness, exactness, evoke, and the (harmonic mean) F measure. Presentation of the incorporated Spectral Clustering and Random Forest loom is found to provide important developments in all the metrics listed, contrasted to a satisfied filtering technique such as LDA joined with text message deletion done arbitrarily or in an adaptive fashion using adversarial learning. The Spectral Clustering approach is strong against the lack of content.

Gowtham et.al [2]: Proposed a study, the features of legitimate and phishing webpages were examined in depth, and support on this analysis, this approach proposed heuristics to take out 15 characters from such webpages. These heuristic results were fed as an contribution to a trained machine learning algorithm to identify phishing sites.To the webpages before alarmed heuristics, this approach worn two initial screening modules in this system. The first component, the preapproved site identifier, verify webpages against a confidential of white-list maintained by the user, and the second part, the Login Form Finder, categorize webpages as legitimate when there are no login appearances present.

Deshmukh et al.[11]: Proposed a approach as cyber crime is technology based fault committed by technocrats. This paper deals with modification of cyber crime like Packet Sniffing, Salami Attack, Bot Networks and Tempest Attacks. It also contains real world cyber crime suitcases their situation and modus operandi. The worldwide malware, rate spam rate and phishing rate is rising speedily. And there is a latent shock of cyber crime on consumer trust, economics and production time. The contradict ways similar to Intrusion Detection, GPRS Security architecture and Agent Based Distributed Intrusion Detection System and prevention System are utilized for safety reason.

Verma et al.[13]: Propose a approach that merged statistical examination of website URLs with machine learning methods will give a additional precise classification of phishing URLs. Employing a two-sample Kolmogorov-Smirnov examination along with other description. Thus, correctness of phishing URL categorization can be very much improved through the use of these statistical measures.

PROPOSED METHODOLOGY

The proposed problem aims to study the phishing detection by using web logo approach.The methodology comprises of following two processes:

First process will capture the screenshot and perform the approach directly to extract the logo. This approach has a few advantages. As the research work will focus on as a replacement for finding the logo image from a pool of downloaded images (image income of a query webpage),Will be capture the screenshot and directly extract the logo.This approach has a few advantages. As, the captured screenshot is actual offer the web content, which means there is no other secret image. By directly capturing the screenshot will provide the actual web content which is usually used to optimise website loading speed.Google image search will provide the undesired result by using sprite type of images as a query result even through the logo existed within the sprite image. Another advantage is the logo removal from the poster image of a website will be more precise. In other words,by directly extracting the logo images will provide no other non-logo images.Second process will utilize the Google image search to retrieve the actual identity. As the link between the logo and domain name is special, it is realistic to treat the domain name as the identity. As a contrast between the domain name returned by Google with the one from the query website will allow us to differentiate a phishing website from a legitimate website .Using a graphical element such as a logo to detect a phishing website. alexa rank of the website is extracted and matched under the range less than 10000 for providing more accuracy in phishing detection.

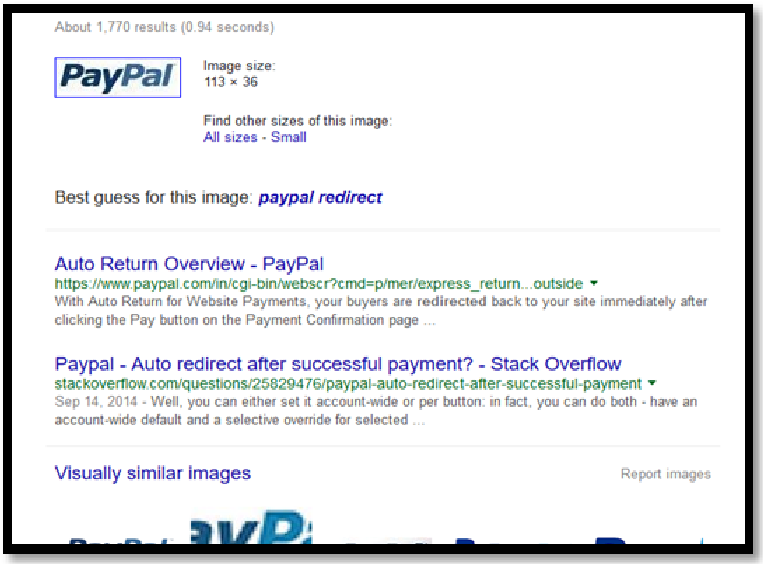
Consistent identity indicates a legitimate website and inconsistent identity indicates a phishing website.

A. Design Consideration

In the design consideration,the proposed approach considered the structure of data flow for the design of the experimental setup. We started by analysing the requirements. The requirements can be listed as follow:

1. Database of Phishing and Non-Phishing websites.
2. Screenshots of the Website under consideration.
3. Processing Tool to Extract Website Logo Image.
4. Automated tool to Detect the Phishing Site by Logo Image Screenshot.

To verify these design considerations, we started by collecting the database of phisihing and non-phishing sites. Phishload [5] is an open source database that have been used in the work. Screenshots of the webpages are collected from PhishTank [6]. A url from PhishTank returns the screenshot of the webpage if available. Processing Tool to extract website logo image is developed in Java. It is an assisted cropping tool and the user has to draw a rectangle around the logo image and the image is extracted. The detection of Phishing Site is done by another tool developed in java.



About 1,770 results (0.94 seconds)

Image size:
113 × 36

Find other sizes of this image:
All sizes - Small

Best guess for this image: *paypal redirect*

Auto Return Overview - PayPal
https://www.paypal.com/in/cgi-bin/webscr?cmd=p/mer/express_return...outside
With Auto Return for Website Payments, your buyers are redirected back to your site immediately after clicking the Pay button on the Payment Confirmation page ...

Paypal - Auto redirect after successful payment? - Stack Overflow
stackoverflow.com/questions/25829476/paypal-auto-redirect-after-successful-payment
Sep 14, 2014 - Well, you can either set it account-wide or per button: in fact, you can do both - have an account-wide default and a selective override for selected ...

Visually similar images

Report images

Fig.1. A Google Image Search Result

In figure1 shows that after searing the logo image of a website in the google image search,it shows the result that whether it is a phishing website or a legitimate website.It shows the best guess of the searched logo image.

B. Flowchart Of Concept

The flowchart of the concept is mentioned in the figure 2. As per the flow of code, we will load a phishing url (known to us only). We will load an phishtank id from database and we will perform a search to get its screenshot image. After getting the screenshot image, we will open the Crop Image Tool and Crop the Logo Region. In each step, a logo image is taken from the database which is cropped and stored in a database. The image is uploaded to google image search website. The google returns the results in terms of a best guess value and some number of urls (search results of websites). If the query logo's URL exists in the list of urls returned by google image search, the website is marked as legitimate, instead, if the website is not directly listed in the set of urls in the list, the alexa rank of the website is extracted and matched under the range less than 10000.

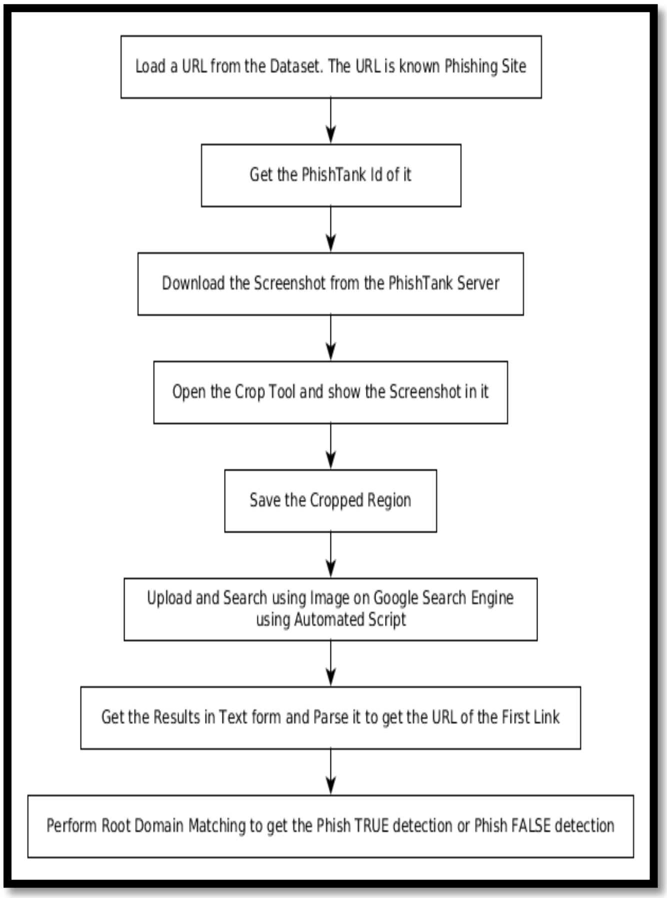
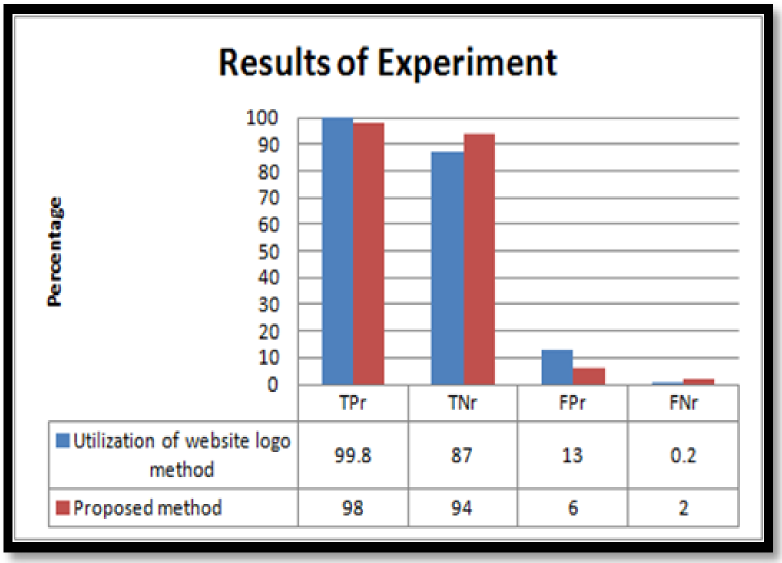

Load a URL from the Dataset. The URL is known Phishing Site
Get the PhishTank Id of it
Download the Screenshot from the PhishTank Server
Open the Crop Tool and show the Screenshot in it
Save the Cropped Region
Upload and Search using Image on Google Search Engine using Automated Script
Get the Results in Text form and Parse it to get the URL of the First Link
Perform Root Domain Matching to get the Phish TRUE detection or Phish FALSE detection

Fig.2: Flowchart of the Proposed Work

C. Results and Discussions

In each step, a logo image is taken from the database which is cropped and stored in a database. The image is uploaded to google image search website. The google returns the results in terms of a best guess value and some number of urls (search results of websites). If the query logo’s URL exists in the list of urls returned by google image search, the website is marked as legitimate, instead, if the website is not directly listed in the set of urls in the list, the alexa rank of the website is extracted and matched under the range less than 10000.



	TPr	TNr	FPr	FNr
🟦 Utilization of website logo method	99.8	87	13	0.2
🟥 Proposed method	98	94	6	2

Fig.3: Graph of Performance Analysis of Our System compared to Utilization of website logo based Method

In Figure3 the True Positive rate of our system is 98% while the True Positive rate of Utilization of logo image based mehod is 99.8 %, further, the True Negative rate of our system is 94% whereas the true negative rate of Utilization of logo image based mehod is 87 %.

Conclusion

In this work we have developed a new method to detect phishing websites based on the logo image and the base url of the website. The system has shown a 98% detection rate of phishing website becuase the logo used by phishing website returns the search of original websites or other websites that have backlinked the original website but the test website's url never appears in the search result. Thus making a 98% accurate detection because some of the alexa ranks were skewed towards base URL. The False Positive Rate is imporved from previous work by more than 50 %. But since many websites metion the same logo image to backlink a popular website, the masking effect happens and thus real websites are detected as phishing website too.

Future Scope

In the future work, we can add more parameters like Google PageRank, number of backlinks etc in order to increase the overall confidence towards phishing as well as non-phishing website.

REFERENCES

[1] Chiew, K.L., Chang, E.H. and Tiong, W.K., 2015. Utilisation of website logo for phishing detection. *Computers & Security, 54,* pp.16-26.

[2] Gowtham, R. and Krishnamurthi, I., 2014. A comprehensive and efficacious architecture for detecting phishing webpages. *Computers & Security*, *40*, pp.23-37.

[3] DeBarr, D., Ramanathan, V. and Wechsler, H., 2013, June. Phishing detection using traffic behavior, spectral clustering, and random forests. In*Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on* (pp. 67-72). IEEE.

[4] Tout, H. and Hafner, W., 2009, August. Phishpin: An identity-based anti-phishing approach. In *Computational Science and Engineering, 2009. CSE'09. International Conference on* (Vol. 3, pp. 347-352). IEEE.

[5] Phishload. 2016. *Phishload*. [ONLINE] Available at: http://www.medien.ifi.lmu.de/team/max.maurer/files/phishload. [Accessed 01 July 2016].

[6] PhishTank | Join the fight against phishing. 2016. *PhishTank | Join the fight against phishing*. [ONLINE] Available at: http://www. K. L. Chiew, E. H. Chang, S. N. Sze, and W. K. Tiong, “Utilisation of website logo for phishing detection,” *Comput. Secur.*, pp. 1–11, 2015.

[7] J. Hong and L. Cranor, “CANTINA : A Content-Based Approach to Detecting Phishing Web Sites,” pp. 639–648, 2007.

[8] J. Lee, D. Kim, and L. Chang-Hoon, “Heuristic-based Approach for Phishing Site Detection Using URL Features,” *Adv. Comput. Electron. Electr. Technol.*, pp. 131–135, 2015.

[9] Cao Y, Han W, Le Y, "Anti-phishing based on automated individual white-list," Proceedings of the 4th Workshop on Digital Identity Management., pp. 51e60,2008.

[10] Prakash P, Kumar M, Kompella RR, Gupta M. PhishNet, "predictive blacklisting to detect phishing attacks," INFOCOM 2010 29th IEEE International Conference on Computer Communications., pp. 346e50,2010.

[11] Deshmukh, J.J. and Chaudhari, S.R., 2014. Cyber crime in indian scenario–a literature snapshot. *International Journal of Conceptions on Computing and Information Technology*, *2*(2).

[12] Zhuang, W., Jiang, Q. and Xiong, T., 2012, June. An intelligent anti-phishing strategy model for phishing website detection. In *Distributed Computing Systems Workshops (ICDCSW), 2012 32nd International Conference on* (pp. 51-56). IEEE.

[13] Verma, R. and Dyer, K., 2015, March. On the Character of Phishing URLs: Accurate and Robust Statistical Learning Classifiers. In *Proceedings of the 5th ACM Conference on Data and Application Security and Privacy* (pp. 111-122). ACM.

[14] Bian, K., Park, J.M., Hsiao, M.S., Belanger, F. and Hiller, J., 2009, July. Evaluation of online resources in assisting phishing detection. In*Applications and the Internet, 2009. SAINT'09. Ninth Annual International Symposium on* (pp. 30-36). IEEE.

[15] Ali, M.M. and Rajamani, L., 2012, March. Deceptive phishing detection system: from audio and text messages in instant messengers using data mining approach. In *Pattern Recognition, Informatics and Medical Engineering (PRIME), 2012 International Conference on* (pp. 458-465). IEEE.

[16] Tan, C.L. and Chiew, K.L., 2014, December. Phishing website detection using URL-assisted brand name weighting system. In *Intelligent Signal Processing and Communication Systems (ISPACS), 2014 International Symposium on* (pp. 054-059). IEEE.

[17] Fang, X., Koceja, N., Zhan, J., Dozier, G. and Dipankar, D., 2012, June. An artificial immune system for phishing detection. In *Evolutionary Computation (CEC), 2012 IEEE Congress on* (pp. 1-7). IEEE.

[18] Nguyen, L.A.T., To, B.L., Nguyen, H.K. and Nguyen, M.H., 2014, October. An efficient approach for phishing detection using single-layer neural network. In *Advanced Technologies for Communications (ATC), 2014 International Conference on* (pp. 435-440). IEEE.

[19] Jo, I., Jung, E.E. and Yeom, H.Y., 2010, August. You're Not Who You Claim to Be: Website Identity Check for Phishing Detection. In *Computer Communications and Networks (ICCCN), 2010 Proceedings of 19th International* phishtank.com/. [Accessed 01 July 2016].

[20] Anti-Phishing Working Group, 2014. Phishing Activities Trends Report. *avail-able at http://docs. apwg. org/reports/apwg_trends_report_q1_2014. pdf*.

[21] Satane, V.V. and Dasgupta, A., 2013. Survey Paper on Phishing Detection: Identification of Malicious URL Using Bayesian Classification on Social Network Sites. *International Journal of Science and Research (IJSR)*.

[22] Khonji, M., Iraqi, Y. and Jones, A., 2013. Phishing detection: a literature survey. *IEEE Communications Surveys & Tutorials*, *15*(4), pp.2091-2121.

[23] Afroz, S. and Greenstadt, R., 2011, September. Phishzoo: Detecting phishing websites by looking at them. In *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on* (pp. 368-375). IEEE.