Design Approach to Recognize Phishing Website by a Combination of Two Methods Using the Advantage of IP Address and Webpage Characteristics

Article · May 2019

CITATIONS
0

READS
335

5 authors, including:

Muhanad Mohammed
Thi Qar University
5 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Bahaa Hussein TAHER Ghrabat
University of Sumer
12 PUBLICATIONS **213** CITATIONS
SEE PROFILE

Raad Ghazi Al-Azawi
University of Babylon
5 PUBLICATIONS **13** CITATIONS
SEE PROFILE

Design Approach to Recognize Phishing Website by a Combination of Two Methods Using the Advantage of IP Address and Webpage Characteristics

Muhanad Mohammed Kadum, Computer Science Department, College of Computer Science and Mathematics, University of ThiQar, Iraq. E-mail: muhanad@utq.edu.iq

Bahaa Hussein Taher, Computer Science Department, College of Computer Science and Information Technology, University of Sumer, Iraq. E-mail: ghrabiuk@gmail.com

Raad Ghazi Al-Azawi, Information Technology Department, College of Information Technology, University of Babylon, Iraq. E-mail: rgazi44@gmail.com

Abstract--- Cybercrime is one of the challenges of the 21st century.One of the cybercrimes is phishing attacks and one of them is Phishing Websites. There are thousands of phishing website, which aims to bluffer users and stealing important information, Such as account information and banking information. In this paper, we proposed a combination of two methods to detect website phishing. This first method is to make a list (MVWL) of popular websites and their IP addresses. However, the used method focuses on comparing a screenshot taken for both the popular website as well as the Doubtful one. While the second method focuses on analyzing website features to find out if, the website trusted or not. Where, the features have been selected in this method deduce from reviewing other researcher's works, as well as our studying for the data set that we examined.

Key words--- Phishing Attack, Phishing Website, Suspicious Website, Doubtful Website and MVWL (Most Visited Website List).

I. Introduction

Phishing attack is one of cybercrime, which carrying out by many ways, one of these is social engineering Technologies, to deceive Internet users to reveal personal and confidential information[1]. The phishers looking to gain victims data such as bank accounts, user name and passwords, the methods used are miscellaneous. So it can bypass the existing anti-phishing techniques[2] (E.g. Blacklist and Whitelist, Decision Trees and Rule Induction)[3]. In addition, the educated user may have experience and, at times, be vulnerable to attack [4]. In this type of attack, the attacker creates a fake web page by copying or making a minor change to the legitimate page, so that the Internet user cannot distinguish between phishing and legitimate web pages.

Phishing websites often contain an equivalent visual plans for the genuine sites particularly the visual style [5]. Moreover, the most imperative component seen by the greatest number of clients will be the target. A phishing site does not give comparative administration to the relating real site. An assailant could download any substantial website page to make a phishing page. A phishing site may contain a few connections that divert clients to the comparing real site (for instance, if the client finds any troubles in getting to their record/account, tap on the Help interface , then the site sidetracks to the Help area of the Legitimately Directed Website To check the hyperlink relationship).

Themost targeted for this type of attacks,as it appear in "Malcovery" reported that in the last quarter of 2013, the five major target companies phishers were Facebook, WhatsApp, UPS, Fargo and Companies House (UK) [6]. "Sheng et al". They showed that women were more likely to be victims of phishing than were men. The same applies to people aged 18 to 25, probably due to a lack of awareness of phishing threats [4, 7, 8]. According to RSA's monthly online fraud report[9].

In this paper, will focus on how to recognize phishing websites by using two deferent methods, the first one, will discussif the website is popular (Alexa website)1, so we create a list of most visited website(MVWL), this list contend the name as well as the IP address for each website from top 500 website. When the doubtful website (DW) is in the (MVWL), then will get the IP address of the DW and, comparing it with original website's IP address,

Received: 20 September 2018/Accepted: 15 October 2018

instead of comparing an image for both websites[10]. Furthermore, when the suspicious website is not in our list, then its characteristic will test and this will represent the second method.

We inspected in excess of 13,000 phishing websites taken from Phish Tank2 and found that 4170 of it was in the most visited list, so we use the first approach to check them. Weusing the second approach for the others. Furthermore; 2750 Unknown websites also taken from Phish Tank, has been examined. This paper will organize as follow: section2 Related works, section3 proposed methods, section4 results, evaluation, and section5 Conclusion.

II. Related Works

The most imitated entities are the foremost visited web content on the Internet from that Associate in wrongdoer will port (like paypal.com, ebay.com or facebook.com, etc.). List of the foremost visited websites is obtainable on alexa.com. This phishing methodology often detected by a comparison of the screenshot of the detected web site to alternative screenshots of the foremost visited web content. This can be a correct methodology to discover a phishing website [11]. There are issues that must be solve like ads, animations and time-dependent transmission. This may be solved by more screenshots that are taken at different times once loading the page. Another methodology to discover the identical content of the web site is taking all the text from the detected web site and comparison it to the foremost visited web content. If there is a high share of conformity, the website is phishing.

The Anti-Phishing social unit (APWG) reportable that they determined a lot of phishing attacks within the half-moon of 2016 than in any different 3 month amount since 2004 once they started aggregation information. In addition, APWG reportable the variety of distinctive phishing websites detected hyperbolic 250 % between October 2015 and March 2016 [12].

“Ponemom” Institute estimates annual losses of phishing attacks for a particular company to be $3.7 million, which forms 48% of losses, related to prices from loss of worker productivity [13].Spear phishing attacks that are winning at corporations and establishments like Target, Sony, and even the Pentagon and White House, price on average around $1.8 million per incident [14].

Despite the tremendous rate of growth, “Vishwanath” and colleagues [15] purpose out that the prevalence of phishing attacks diminish client confidence and trust in online commerce and communication, ensuing in hyperbolic operational prices for online retailers. Thus, analysis that focuses on however to defend these varieties of attacks is a high priority not solely for researchers however conjointly for IT practitioners.

III. Proposed Methods

Since most researchers are focusing on detecting phishing websites either on source code analysis or relying on the URL, while others are depending on comparing screen shoots of suspicious websites with the original trusted websites.

The proposed system, suggest that the examination of websites depending on two criteria, the first, depending on the list of most visited websites as shown in figure1, and the second depending on the characteristics of the website itself as shown in table 1 and figure 2.

Consequently, our work will divide in to two phases. At phase, one a list will be create including the most visited websites in the world, in addition to the IP address of each website. This list will depend on obtained (from data set) from (Alexa website), where those websites almost trusted.

This list will call MVWL (Most Visited Website List), will be daily updating when Alexa update their list of top websites.

Python Software has been developed, in addition using other scripts which also created, to enable us efficiently using Linux commands like (how is, dig, host and nslookup), to get the websites information just like (IP address, Doman name expiration , age and so on). Then used in the python application with needed data to establish our goal. Furthermore, all the work in phase one and two implemented using Ubuntu 16.10 environment.

To check any doubtful website (DW) it will be firstly test if it’s in the MVWL or not, if not the test will change to phase two, if yes then the IP address of the checking website will obtain. After that this IP should be compare with the IP’s in the MVWL, if the DW IPs belongs to MVWL then it is a trusted website else, its phishing website, figure 1 illustrate that. While the following steps will describing our full work.

Received: 20 September 2018/Accepted: 15 October 2018

Step1: Begin.

Step2: Create the MVWL, and read the DW;

Check if DW belongs to MVWL; if true get DW IP address and check if it's belongs to MVWL.

Step3: Check the URL characteristics.

Step4: Check the Doman name characteristics.

Step5: Check the Source code characteristics.

Step6: end.

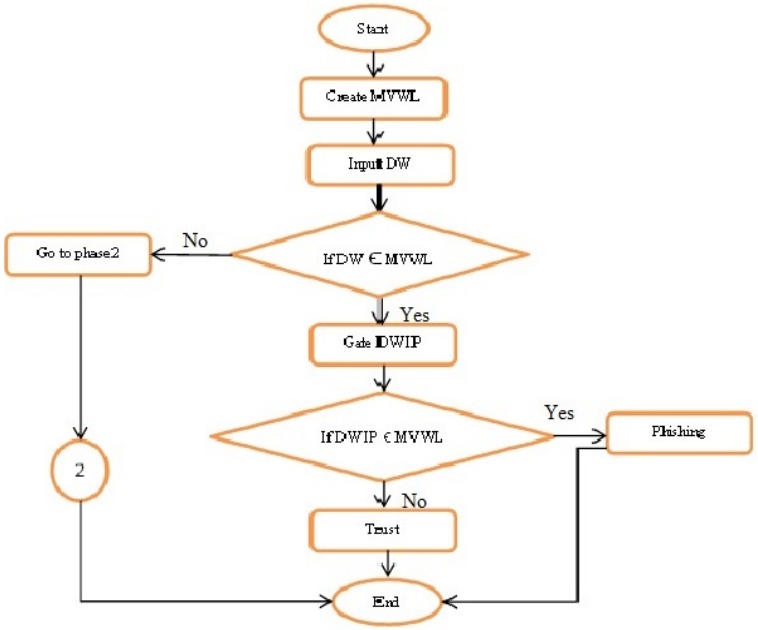

Fig. 1: Phase One Steps

On phase two, the web page characteristics will be examine to find if the DW is trusted or not. The characteristics that we will adopt to know the reliability of the websites will be classify into the properties of the URL, Domain name and the characteristics of the source code, table1 will summarizes them. For this the phase will consists of three levels of testing.

Table 1: Website Characteristics Examined in Phase Two

URL characteristics	Doman name characteristics	Source code characteristics
IP in URL	using (−) symbol in Doman name	source code seen enabled
@ symbol in URL	Include https in Doman name	Using <iframe> tag
Using https protocol	Doman expires	Using java script with no link
Using certificate	Doman age	There is an HTML file as a downloadable page
DOM based XSS	No. of page in website	'**target**' attribute with "**_blank"**
		Using "mail()" or "mailto:" Function to Submit User Information

In level 1, any questionable website that has been converted from the previous phase will be tested based on the characteristics listed in Table 1, as the test will start with the URL part.Where, if there is an IP address with both V4 and V6 in the URL, so it conceder a phishing website. If there is a (@) symbol in URL, also it will be a phishing website. Furthermore; if (DOM based XSS) used in URL so it's phishing site, the following example will explain this type of fishy,(http://site.org /page.html? variable =< s c r i p t >doEvilCode ( ); </ script >) where the EvilCode and script talking in URL used to execute those malicious code directly in browser. In addition to that, if the URL has the https protocol and certificate is testing, so if the website has the both, then it will considered a trust website.Else the test will change directly to Doman level, because from our study, we find that many trusted website

Received: 20 September 2018/Accepted: 15 October 2018

don't use them, so it will be considered a week points for the website has been test and if it's meet any of other characteristics in table1 so this website will consider a phishing one. Our application, checking the above URL features very fast and makes a decision if this site is phishing, trust or goes to the next level of testing.

In level 2, the Doman name part will examine as a part of URL, As well as if there is (https) appears in URL after (www.) then this website is considered a phishing one.Else If there is a (−) symbol in Doman name, then the DW is phishing. Else our application will check the Doman name expiration(J)and age (K) by using (howis) command, so if the expirer is less than 5 months and the age is less than 6 months then the website page number (N) will checking, so if the page No. is less than 100 page then the website is phishing site.Else the software will change to level 3. The following function will describe this feature.

$$\text{If } J<5 \wedge K<6 \wedge N<100 \begin{cases} \text{True>> phishing} \\ \text{False go to level 3} \end{cases}$$

It should be note here that the number mentioned (5months, 6 months and 100 pages) have been drawn through the study conducted on the data set that we examined.

In level 3, the DW source code will examine, firstly the ability of seen the source code is checked, where if feature is available will continue with other checking steps, if not the DW is phishing website. Furthermore, the excessive usage of inline frames <iframe> tag should be teste, where it commonly use to feed in the web page from external links with 'src" attributes where the syntax of it as (<iframesrc="URL" width="100px" height="100px">).In this case we have to test the link (URL) has given iframe tag after (src=), so our platform will checking the link starting from phase one, and from the study conducted on the data set obtained from Phishing Tank for more than 13,000 website classified as valid phishes for eight months starting from April 2018 tile December 2018. We found that the web site that use the <iframe> tag with 'src" attributes for more than three times for different links are phishing sites. Also, we found that a high percentage of phishers are included in their codes one of the following features (java script with no link, an HTML file as a downloadable page, 'target' attribute with "_blank" and using "mail()" or "mailto:" Function to Submit User Information).For that our platform has been prepared to consider any web site contain any of the previous features as phishing web site. Figure 2 illustrate the procedure followed in phase two.

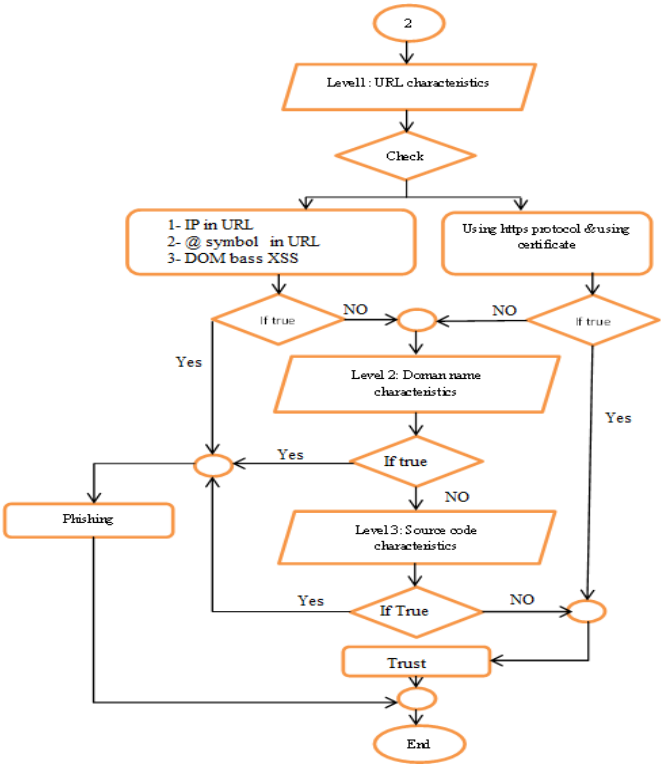

Fig. 2: Phase Two Steps

Received: 20 September 2018/Accepted: 15 October 2018

IV. Results and Evaluation

When the proposed method has been applied on the data set that we collected before, the outcome obtained and it will indicate in the following tables according to the stages has been listed above, were table2 will shows the results got from phase one. Table3 will illustrate the results of phase two divided in to three levels depending on the reparation of websites that emerged according to each test in addition to what it represents as a percentage of the total number of websites that have been test.While the total number of valid phishing websites has been examined was 13,000 as we mentioned before, and another 2750 Unknown websites obtained from phish thank has been checked in our application. Furthermore, table 4 will describe the overall usage of the application in general.

Table 2: Shown the Result Obtain from Phase One

Phase one					
Valid phishes websites examined			Unknown websites examined		
Total number	No. of captured as phishing	No. of website transfer to phase two	Total number	No. of captured as phishing	No. of website transfer to phase two
13000	4170	8830	2750	12	2738

By applying our application, in phase one we can notice that there are about 4170 websites from valid fishes has been examine were fake websites and the phishers trying to inspire users on the internet by creating websites similar to most visited websites. As well as there are 12 website are Unknown websites, Also they were forged to similar of the most visited sites on the net, and the fact that the application discovered it by comparing the IP addresses of the suspicious sites with the original URLs as shown in Figure5 and Table 2, which is a positive advantage of our application .

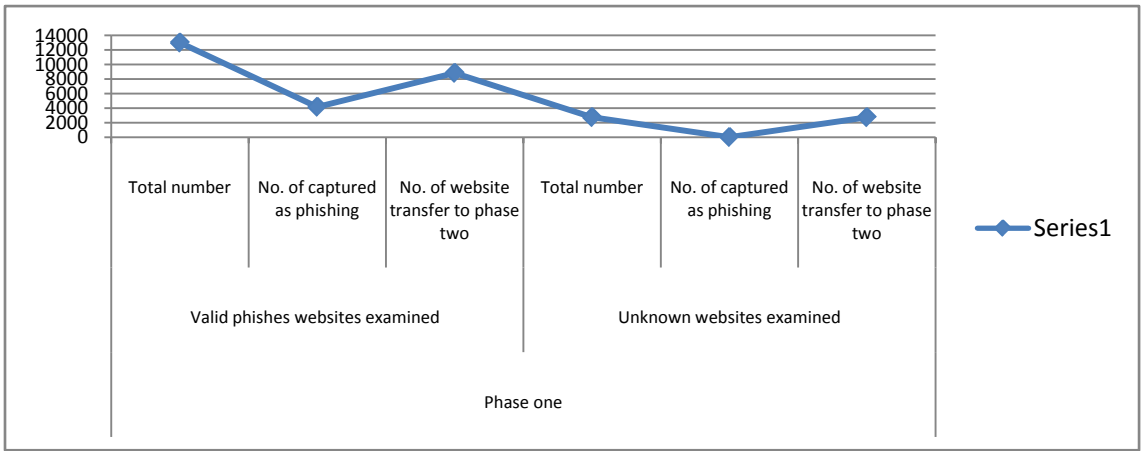

Fig. 3: Evaluation of Phase One Testing

While, there are about 8830 website from the Valid phishes and 2738 from the Unknown ware passed to the phase two, and the result of their examination shown in table3, while the evaluation will illustrate in figure 4.

Table 3: Shown the Result Obtain from Phase Two

Phase two				
Characteristic	8830 Valid phishes websites		2738 Unknown websites	
	No. of captured as phishing	Percentage of total website	No. of captured as phishing	Percentage of total website
Level 1				
IP in URL	21	0.24%	Non	——
@ symbol in URL	96	1.08%	Non	——
DOM based XSS	367	4.156%	4	0.146%
Using https protocol and certificate Where those are trust websites	Non	——	Non	——
Level 2				
using (−) symbol in Doman name	169	1.91%	Non	——
Include https in Doman name	74	0.838%	Non	——
If J<5 ∧ K<6 ∧ N<100	1983	22.457%	47	1.71%
Level 3				
source code seen enabled	10	0.11%	Non	——
Using <iframe> tag	567	6.42%	1	0.036%
Using java script with no link	584	6.61%	Non	——
There is an HTML file as a downloadable page	53	0.6%	Non	——
target' attribute with "**_blank"**	1826	20.6%	2	0.07%
Using "mail()" or "mailto:" Function to Submit User Information	2437	27.599%	61	2.23%

Received: 20 September 2018/Accepted: 15 October 2018

Tables (2 and 3) above indicate that, there was some website from the valid phishes data set not appear as a phishing website in our application. Either for the other data set the Unknown, also in this phase some of those websites classified as a phishing websites, where this gives our application anther points if we compare it with the anti-phishing applications used, as the other application and researchers don't recognize them as phishing websites and they still classify them as Unknown websites in phish tank. As well as figure 6 illustrate phase two evaluation.

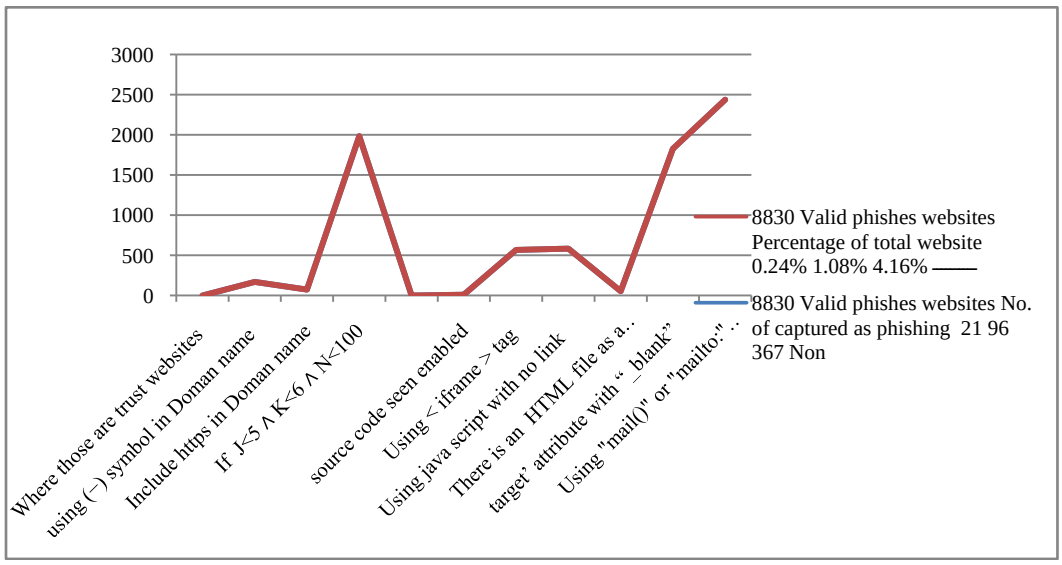

Fig. 4: Evaluation of Phase Two

Table4 will show the final of results for using the application on 13000 valid phishes and 2750 Unknown websites.

Table 4: Final of Results of theApplication

Final Results			
13000 Valid Phishes websites tested		2750 Unknown websites tested	
Total No. of captured as phishing	Percentage	Total No. of captured as phishing	Percentage
12357	95.05%	127	4.61%

As shown in table 5, the final results of using the proposed method and the application has been made according to this method, and after checking 13000 Valid phishes websites taken from phish tank, the percentage of detecting the phishing websites was 95.05%.Add to that our method successfully to recognize 4.61% from 2750 Unknown website as a phishing websites. It should also be noted that the application, which was designed using Python 3 and bash script is not consumed the processor and RAM and that was what make us chosen those programing languages to build our application In addition to using the ability of python in different operating systems like (Linux, Mac, and Windows). While table 5 shown a comparison for our approach results with other researcher results.

Table 5: Results Evaluation and Comparison

Method	Data set used from	Efficiency
Liu et al. [16]	Phish Tank	92.1%
Fields and Latent Dirichlet Allocation (LDA)[17]	Phish Tank	88.1%
Novel techniques for detecting phishing sites and their targets[18]	Phish Tank	90.5-97.3%
Our approach	Phish Tank	95.05%

V. Conclusion

The wide spread of phishing sites and the frequent occurrence of internet users as victims is what led us to engage in this type of research. Through this study, we conducted during the past months and from April 2018 on more than 13000 sites registered as a site of phishing. We suggest this method of detecting these websites to warn users and enable them to distinguish whether they are trust websites or phishing one. Based on this method we have designed the application using the languages of Python and Bash Script to achieve the goal of the search.

Received: 20 September 2018/Accepted: 15 October 2018

The results were presented in the results section, were the efficiency of application was 95.05%, which is a very good result compared with the rest of the researches in this field. This evidenced by the ability of the application to distinguish some sites that did not classified by the other researchers and developers as phishing sites.

This accomplished by testing a data set of website classified as unknown from the phish tank. After working in the Linux environment on the (Ubuntu 16.10) system, in the future, we will going to work on other operating systems and this is what made us use the Python language as the basis for the application.

References

[1] Almomani,A., Gupta, B., AtawnehS. *et al.*, A survey of phishing email filtering techniques.*IEEE communications surveys and tutorials* **15** (4) (2013) 2070-2090.

[2] Mishra, A. andGupta, B.B. Hybrid solution to detect and filter zero-day phishing attacks. In *Proceedings of the Second International Conference on Emerging Research in Computing, Information, Communication and Applications*, 2014, 373-379.

[3] Qabajeh, I., Thabtah, F. and Chiclana, F. A recent review of conventional vs. automated cybersecurity anti-phishing techniques. *Computer Science Review***29** (2018) 44-55.

[4] Sheng, S., Magnien, B., Kumaraguru, P., Acquisti, A., Cranor, L.F., Hong, J. and Nunge, E.Anti-phishing phil, the design and evaluation of a game that teaches people not to fall for phish. In *Proceedings of the 3rd symposium on Usable privacy and security*, ACM, 2007, 88-99.

[5] Jain, A.K.and Gupta, B.B. A novel approach to protect against phishing attacks at client side using auto-updated white-list. *EURASIP Journal on Information Security*(1)2016.

[6] Steve Ragan, *Senior executives blamed for a majority of undisclosed security incidents*.(2013).

[7] Sheng, S., Holbrook, M., Kumaraguru, P., Cranor, L.F. and Downs, J. Who falls for phish, a demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2010, 373-382.

[8] Arachchilage, N.A.G., Love,S. and Beznosov, K.Phishing threat avoidance behaviour. An empirical investigation. *Computers in Human Behavior* **60** (2016) 185-197.

[9] Gupta,B. Arachchilage, N.A.and Psannis,K.E. Defending against phishing attacks, taxonomy of methods, current issues and future directions. *Telecommunication Systems* **67**(2) (2018) 247-267.

[10] Corona, I., Biggio, B., Contini, M., Piras, L., Corda, R., Mereu, M. andRoli, F. Deltaphish, Detecting phishing webpages in compromised websites. In *European Symposium on Research in Computer Security*, Springer, Cham. (2017) 370-388.

[11] Adebowale,M., Lwin, K., Sánchez,E. *et al.*, Intelligent web-phishing detection and protection scheme using integrated features of Images, frames and text. *Expert Systems with Applications* (2018).

[12] Jain, A.K. and Gupta, B.B. PHISH-SAFE, URL features-based phishing detection system using machine learning. In *Cyber Security, Proceedings of CSI*, Springer Singapore, 2018, 467-474.

[13] Le, D.N., Kumar, R., Nguyen,G.N. *et al.*, *Cloud Computing and Virtualization*, John Wiley and Sons (2018).

[14] Thomas, J. Individual cyber security,*Empowering employees to resist spear phishing to prevent identity theft and ransomware attacks* (2018).

[15] Canfield, C.I.and Fischhoff, B.Setting Priorities in Behavioral Interventions. An Application to Reducing Phishing Risk.*Risk Analysis* **38**(4) (2018)826-838.

[16] Wenyin, L.,Liu,G., Qiu,B. *et al.*, Antiphishing through phishing target discovery. *IEEE Internet Computing* **16** (2) (2012) 52-61.

[17] Ramanathan,V. and Wechsler,H.Phishing detection and impersonated entity discovery using Conditional Random Field and Latent Dirichlet Allocation. *computers and security* **34** (2013) 123-139.

[18] Marchal, S., Saari, K., Singh, N. and Asokan, N. Know your phish, Novel techniques for detecting phishing sites and their targets. In *IEEE 36th International Conference on Distributed Computing Systems (ICDCS)*, IEEE, 2016, 323-333.

Received: 20 September 2018/Accepted: 15 October 2018