Journal of Engineering and Applied Sciences 12 (3): 513-519, 2017
ISSN: 1816-949X
© Medwell Journals, 2017

Websites Phishing Detectio Using URLs Tokens as a Discriminating Features

[1, 2]Ammar Yahya Daeef, [1]R. Badlishah Ahmad and [1]Yasmin Yacob
[1]School of Computer and Communication Engineering,
Universiti Malaysia Perlis (UniMap), Perlis, Malaysia
[2]Middle Technical University, Baghdad, Iraq

Abstract: The Phishing detector must be wide scope to deal with the several strategies used to start the phishing campaign and provides high speed detection to avoid user's unsatisfactionby introducing large delay. Consequently, this word presents wide scope and fast detection system by using URLs tokens as a discriminating features without using any external or content features. The method based on analyzing the percentage of the re-used tokens and the token overlap between phishing and legitimate URLs. This research differs from other research by analyzes URLs collected from different sources and according to, this analysis, a statistical classifier is built and the performance is evaluated to measure the technique effectiveness. The results show that the dictionary of phishing tokens is smaller than the dictionary of legitimate tokens and the token overlap between phishing and legitimate URLs is small. Also, the token overlap rate between different phishing sources is more than compared with legitimate token overlap percentage. The average accuracy of 77% is achieved by this technique.

Key words: Phishing, legitimate, lexical features, URL tokens, statistical classifier

INTRODUCTION

The web has evolved widely in the life of people and since the beginning of Internet in the 1990s a lot of new security issues and threats appear continuously which constitute a challenge to users and security experts as well. Phishing is a cutting edge threat that has an impact on commercial and banking sectors by means of the Internet which delivers huge misfortunes at the level of clients and organizations. Phishing websites have high similitude to the honest ones trying to trap and bait users to enter these websites. In this sort of attack, phishers normally utilize technical and social designing traps together to begin their attacks. The attacks of social engineering are focusing on users not systems intended to get the data of users which are typically touchy and secret (Bozkir and Sezer, 2016).

In spite of the broad field of phishing attack vectors a typical purpose of numerous vectors is the utilization of link misleading victims to phishing websites. Utilization of obfuscated Uniform Resource Locator (URL) and domain names is widely used in phishing attacks (Aaron *et al*., 2014). Anti-Phishing Work Group (APWG) reported that the number of phishing websites increased by 250% in the period from the last three months of 2015 to the first quarter of 2016 as shown in Fig. 1. The total number of discovered unique websites in the first quarter of 2016 is 289,371. Also, steadily rose per month was observed from October 2015 to March 2016 ranged from 48114-123555, respectively (Aaron *et al*., 2014). These statistics demonstrate the significance to distinguish URLs and domain names to battle phishing.

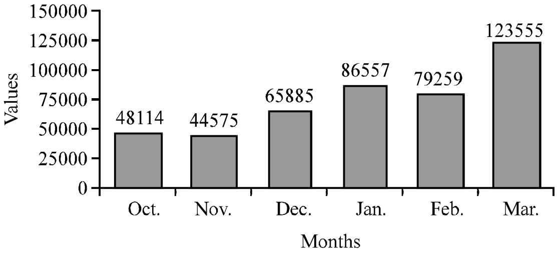

Fig. 1: APWG phishing site trends ist quarter 2016

Most of the research in the field of phishing detection based on website content analysis or use external data from servers to classify URLs as legitimate or phishing class. This research focuses on feature extraction from URLs lexical itself because it needs less processing requirement compared with content or external features. Also, features extraction from URLs lexical can provide wide scope detection depending on the fact that users use URLs directly to search the Internet.

Literature review: A lot of techniques are proposed to detect phishing attacks, most of them based on extracting phishing features either from the website content or using external information. Extract features from website content is resource and time consuming and expose users to

Corresponding Author: Ammar Yahya Daeef, Middle Technical University, Baghdad, Iraq threats by downloading malicious content. Extracting features from external servers website rank, DNS, Whois, etc.) adds more processing time to detect each URL which make such technique not applicable for real time applications.

As an alternative, some methods analysis URLs lexical properties as a discriminating features. Such features are the number of dots in URL, length of tokens and URL length etc. The features extracted by this method are not time consume and prevent downloading malicious code to the user machine. The anatomy of phishing URLs explored by McGrath and Gupta (2008).

Their results state that phishing URLs normally contain the brand name of the target and present different distributions of the alphabet. Also, long URL and short domain name provide strong features of phishing. Take in account this, many research are proposed by utilizing only lexical features extracted from URLs (Blum *et al*., 2010; Khonji *et al*., 2011).

Most of the research (Kan and Thi., 2005; Ma *et al*., 2011) use a bag of word method to represent the lexical features for machine learning classifiers. However, representing lexical features using a bag of word produces high dimension vector which in turn increase the processing time to extract and prepare the features vectors and slow down the training and testing of machine learning classifiers. The authors of PhishStorm (Marchal *et al*., 2014) present URLs lexical analyses in real time. This system is a central classifier placed in front of the email server to detect phishing URLs. PhishStorm uses 12 features extracted by aid of the search engines then these features are fed to machine learning classifier to make the decision. The accuracy achieved by this system is 94.91% combined with a low false positive rate of 1.44%. However, PhishStorm is time consume because of the search engines employed during features extraction process.

The results presented by Khonji *et al*. (2013) analyze the token distribution in both phishing and legitimate URLs. This study confirms that URLs provide additional information than just directing to a resource. Max accuracy achieved from this method is 97%. However, the robustness of this method is not evaluated by training and testing using completely different sources. Finally, some technique uses lexical features combined with different features such who is or DNS information. Such research is found by Thomas *et al*. (2011) this system provides 91% accuracy with 5.54 sec processing time. This high processing time is a result of utilizing external servers to get the host information.

Using complex operations without fully evaluate simpler methods and check the productivity achieved from them is not a good practice. Therefore in this study, we try to analyze URLs lexical features and construct statistical classifier to classify phishing and legitimate URLs. Additionally, we check the robustness of this method by out of sample test using different datasets for training and testing

MATERIALS AND METHODS

This research follows the method proposed in (Khonji *et al*., 2011) to further analyzing URL tokens as a classification features and test the method robustness. The difference between the lexical URL analysis in this research and the one by Khonji *et al*. (2011) is the following:

- This research analyzing token reused percentage and overlap among datasets collected from different sources
- Make sure that each URL is unique in each of the dataset
- Make sure that no repeated host in each datasets
- We tokenized strings using the delimiters specified by Kan and Thi (2005) namely '/', '?', '.', '=', '-' and '_'

The general methodology steps are presented in Fig. 2 which consist of the training and testing phases. Each URL tokenized using the delimiters specified by Ma *et al*. (2011), Kan and Thi, 2005) for example the URL "https://www.paypal.com/my/webapps/mpp/pay-on-ebay" is lexically broken into the following tokens: www, paypal, my, webapps, mpp, pay, on, ebay

All TLDs, (e.g., com, org, edu, est) are removed because they are used commonly in legitimate and phishing URLs as well and therefore indistinctive. To analyze the distribution of tokens in both legitimate and phishing URLs, these URLs are treated one after one to calculate the percentage at which tokens are reused in subsequent URLs. More clearly, the first URL tokens are not seen before then the next URLs appeared to reuse tokens already seen in previous URLs. Java script is written to automate the process described.

A statistical classifier: A binary classifier that constructed to classify each URL in the test phase as either phish or legit class. The classifier is built using a supervised learning phase by extracting the tokens from labeled URLs. After the completion of learning phase, the classifier is fed by unclassified URLs to predict the output class. To predict the output class, each incoming URL is broken into tokens then try to find each token frequency

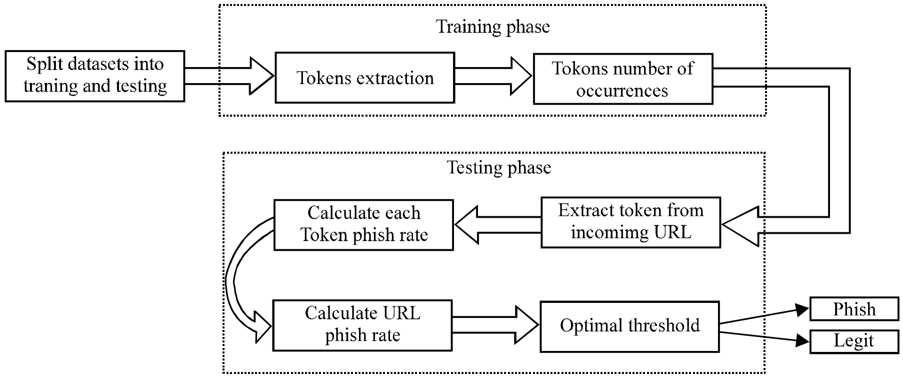
Training phase
Split datasets into traning and testing
Tokens extraction
Tokons number of occurrences
Testing phase
Calculate each Token phish rate
Extract token from incomimg URL
Calculate URL phish rate
Optimal threshold
Phish
Legit

Fig. 2: Research methodology phases

(the number of occurrences) from each class. After that for each token, the phish rate is calculated using Eq. 1:

$$\text{Token phi shrate}_i = \frac{\text{Count}_i \rightarrow \text{Phish}}{\text{Count}_i \rightarrow \text{Phish} + \text{Count}_i \rightarrow \text{Legit}}$$

After calculating the phish rate of each token in the input URL, The phish rate of that URL is calculated by adding the phish rate of all individual tokens and divided by the number of tokens exist in that URL as shown in Eq. 2:

$$\text{URLphishrate} = \frac{\sum_i^N \text{Tokenphishrate}_i}{N}$$

where, N is n is the number of tokens in URL. Each URL in the testing phase is classified as a phish if its phish rate value is more than a certain threshold. The classifier is tested using different values of threshold ranged between 0 and 1 with 0.001 increment for each test.

Datasets: The training data was drawn from four sources: Phishtank.org, Openphish.com, DMOZ.org, and Alexa.com. We collected 20000 phishing URLs from Phishtank and call it Tank dataset. For more closely following the evolving features of phishing URLs and to mimic the real-world scenario, a second batch of 20000 confirmed phishing URLs that were submitted to OpenPhish is collected and call it Open dataset.

To cover the diversity of legitimate websites, our legitimate URLs are gathered from two data sources provided publicly: DMOZ.org and Alexa.com. 20000 randomly chosen non-phishing URLs from DMOZ and we call it DMOZ data set. Also, 20000 randomly chosen non phishing URLs are collected from Alexa and named this dataset as Alexa dataset. Additionally, in order to cover wider URL structures, we also made a list of URLs related to most commonly phished targets (using statistics of top targets from PhishTank and Open Phish) to be part of DMOZ and Alexa datasets.

Table 1: Classifier performance metrics

Evaluation metric	Definition
False Positive Rate (FPR)	The ratio of legitimate URLs misclassified as phishing class divided by the total number of legitimate instances $\text{FRP} = \frac{N_{L\rightarrow P}}{N_{P\rightarrow P} + N_{L\rightarrow P}}$
False Negative Rate (FNR)	The ratio of phishing URLs is classified as legitimate class divided by the total number of phishing instances $\text{FNR} = \frac{N_{P\rightarrow L}}{N_{P\rightarrow P} + N_{P\rightarrow L}}$
True Positive Rate (TPR)	The ratio of phishing URLs classified as phishing class divided by the total number of phishing instances $\text{TPR} = \frac{N_{P\rightarrow P}}{N_{P\rightarrow P} + N_{P\rightarrow L}}$
True Negative Rate (TNR)	The ratio of legitimate URLs classified as legitimate class divided by the total number of legitimate instances $\text{TNR} = \frac{N_{L\rightarrow L}}{N_{L\rightarrow L} + N_{L\rightarrow P}}$
Accuracy	The ratio of correct classification over all attempts of classification $\text{Accuracy} = \frac{N_{L\rightarrow L} + N_{P\rightarrow P}}{N_{L\rightarrow L} + N_{L\rightarrow P} + N_{P\rightarrow P} + N_{P\rightarrow L}}$

Finally, Phish Tank and Openphish datasets are paired with non-phishing URLs from a benign source (either DMOZ or Alexa). We refer to these data sets as the Tank-DMOZ (TD), Tank-Alexa (TA), Open-DMOZ (OD) and Open-Alexa (OA). Figure 3 shows the methodology of datasets merging.

Evaluation Metrics: There are several metrics to measure the quality of binary classification models. We present the most widely used ones that are briefly described in Table 1.

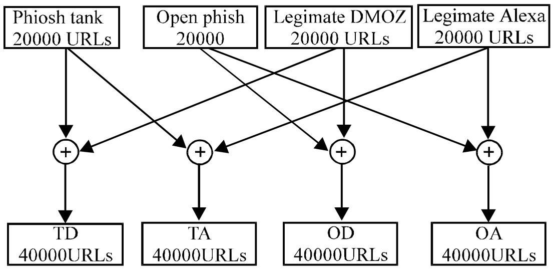

Fig. 3: Dataset merging methodology

RESULTS AND DISCUSSION

The first step of this analysis is the percentage at which tokens are reused in each individual dataset. As shown in Fig. 4, PhishTank dataset has re-usage token percentage reached to 65.26% while the percentage in OpenPhish is 66.85%. Token reuse in legitimate datasets shows less percentage of 46 and 49.98% for DMOZ and Alexa respectively. It is obvious that the percentage of reused phishing URL tokens is higher than legitimate percentage which in turn gives evidence that the dictionary of phishing tokens is smaller than the dictionary of legitimate tokens. Such percentage is logical because of phishers target famous brand frequently and mostly they reuse the same tricks to start the attack in contrast to the huge number of legitimate URLs exist nowadays. Although the dictionary of phishing tokens is less than legitimate one, tokens of legitimate URLs are still predictable as around 50% of the tokens are reappeared or reused. From practical point of view as both classes have limited dictionaries of tokens, this can be exploited to build robust classification model using URLs tokens. To study the common characteristics of the datasets, the token overlap between different sources is explored. As shown in Fig. 5, the overlap between phishing sources is 49.41% which means that even with different sources of phishing URLs, these URLs share big percentage of tokens. This is very motivational point to create robust classifiers. On the other hand, low tokens overlap percentage is observed in legitimate datasets which reached to 15.27%. This is expected because of the wide variety exist in legitimate URLs.

As well as the analysis includes the overlap percentage of tokens between each phishing source and legitimate sources as depicted in Fig. 6. In average, the percentage of tokens overlapping in relation to legitimate and phishing sources around 10%. As a result, the biggest percentage of tokens is not overlapped between phishing and legitimate sources. This observation is very important and promising to build a classification model using tokens only.

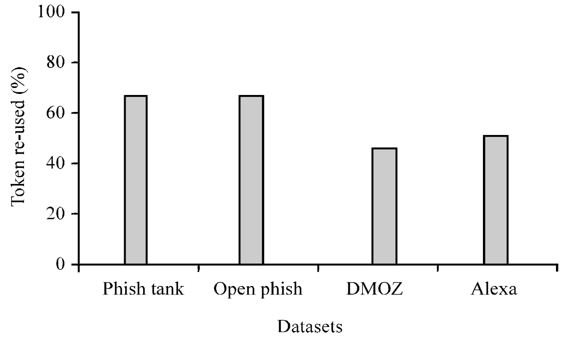

Fig. 4: Percentage of reused tokens in each datasets

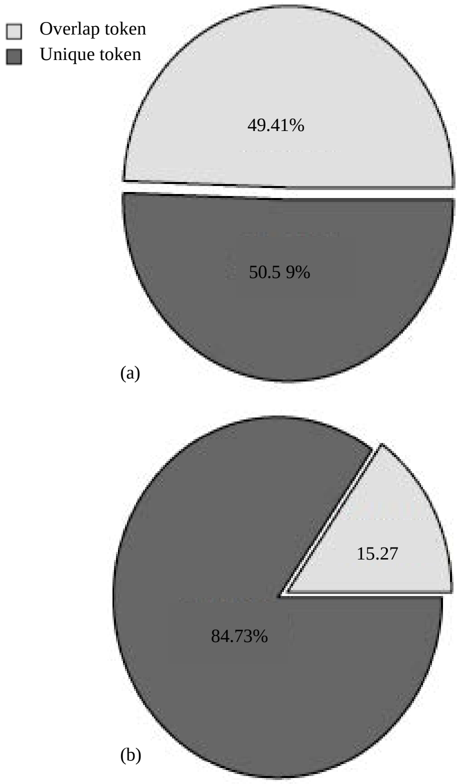

Fig. 5: Overlap percentage of: a) phishing datasets and b) legitimate datasets

The statistical classifier depends on the training dataset to build the classification model then the testing dataset is used to evaluate the generated classifier. As response to that, each dataset is separated into 70% training portion and 30% as testing samples to evaluate the classifier. For each dataset, the optimal threshold is explored by applying thresholds between 0 and 1 with 0.001 increment. The process is repeated for all datasets and optimal threshold is reported according to the maximum accuracy achieved. Figure 7 shows how the

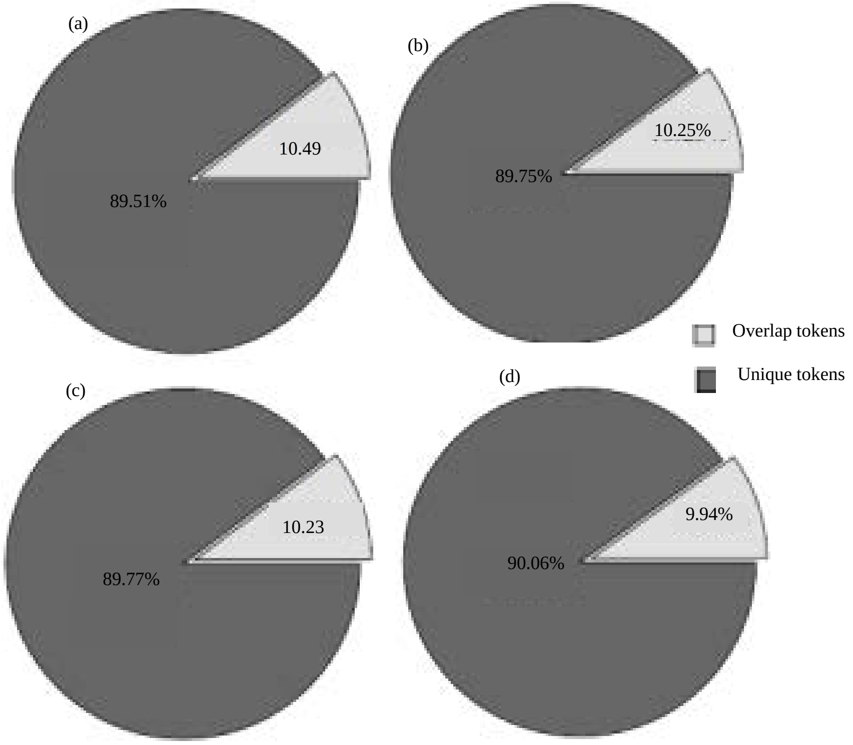

Fig. 6: Overlap percentage between phishing and legitimate datasets: a) PhishTank and DMOZ ;b) PhishTank Alexa; c) Open Phish DMOZ and d) OpenPhish Alexa

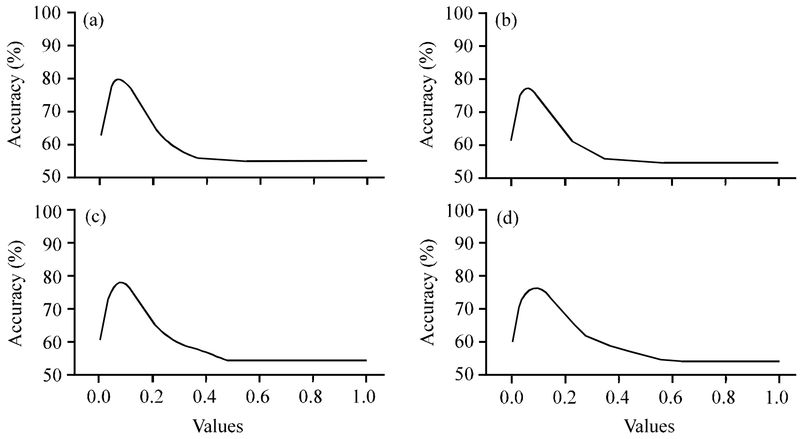

Fig. 7: a-d) Optimum threshold selection for each dataset

classifier accuracy behaves as the threshold is changed. Fig. 8 shows the optimal threshold for each dataset and the corresponding accuracies. The accuracies are not differ significantly with average accuracy 77% because the overlap percentage between the phishing datasets and each of the legitimate URLs source is close to each other.

For a close look at the detailed performance metrics of the statistical classifier on each dataset with optimal threshold, Table 2 presents the results of TPR, TNR, FPR and FNR. The results show that the highest TPR of 86.40% using OA dataset while the highest TNR is 70.99% observed on OD dataset. In general, TPR is higher than TNR on all datasets this is because of the higher percentage at which phishing tokens are reused. Also FPR is higher than FNR which means that more legitimate URLs are miss-classified as phishing class than classify phishing samples as legitimate URLs. Next, the out of

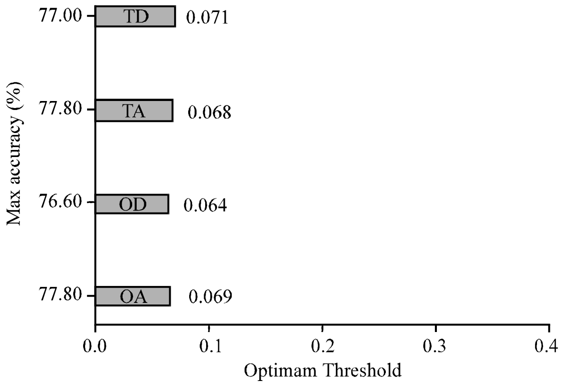

Fig. 8: Optimal threshold and maximum accuracy on each dataset

Table 2: Classifier performance metrics results

Dataset	TPR (%)	FPR (%)	TNR (%)	FNR (%)
TD	84.63	30.47	69.53	15.37
TA	85.32	29.57	70.43	14.68
OD	82.30	29.00	70.99	17.70
OA	86.40	30.90	69.10	13.60

Table 3: Overall rate of errors using mismatched datasets

	Training			
Testing	TD (%)	TA (%)	OD (%)	OA (%)
TD	5.060	7.890	13.00	17.54
TA	7.610	5.020	16.87	12.65
OD	14.71	18.84	5.040	6.780
OA	18.31	14.66	6.520	4.070

sample test based on this method is presented to explore the classifier by training and testing using different datasets.The statistical classifier results of using mismatched datasets for training and testing are shown in Table 3. Based on the results and as expected because of tokens overlap percentage, the error rates are better when training and testing using the same dataset (as shown in the diagonal of Table 3) compared to when mismatched datasets are used for training and testing. When using any combination of phishing and legitimate URLs in the training phase and testing by mismatched phishing URLs only, (e.g., TD, OD) the error rates increased because of the high FN. When the phishing URLs are mismatched and because of the nature of the used classifier, more unseen phishing tokens will be in the testing phase which makes the classifier miss classifying alot of phishing URLs as a legitimate class. The highest error rate observed in this category is 14.71%. In case of legitimate URLs are mismatched only (e.g., OA and OD) in training and testing, the error rates are rising up mostly contributed by FP with the worst value of 7.89%. Finally, when both sources are mismatched, (e.g., TD and OA) this leads to more unseen tokens in testing phase which makes the error rates increased rapidly. The highest error rates are observed in this category with max value reached to 18.84%.

CONCLUSION

This study analyses token distributions in both phishing and legitimate URLs collected from different sources. The results show that the dictionary of phishing tokens is smaller than the dictionary of legitimate tokens. Generally, the token overlap between phishing and legitimate URLs is small. But the overlap rate between different phishing sources is more than compared with legitimate overlap percentage. However, this technique can be effective if the training and testing using the same dataset but in case of out of sample test the error rates increased rapidly. We believe combine this method with high rank lexical features can be the next research step to improve the overall performance.

REFERENCES

Aaron, G., R. Rasmussen and A. Routt, 2014. Global phishing survey: Trends and domain name use in 1H2014. http://docs.apwg.org/reports/APWG_Global_Phishing_Report_1H_2014.pdf.

Blum, A., B. Wardman, T. Solorio and G. Warner, 2010. Lexical feature based phishing URL detection using online learning. Proceedings of the 3rd ACM Workshop on Artificial Intelligence and Security, October 4-8, 2010, Chicago, IL., USA., pp: 54-60.

Bozkir, A.S. and E.A. Sezer, 2016. Use of HOG descriptors in phishing detection. Proceedings of the 4th International Symposium on Digital Forensic and Security (ISDFS), April 25-27, 2016, IEEE, Ankara, Turkey, ISBN:978-1-4673-9865-7, pp: 148-153.

Kan, M.Y. and H.O.N. Thi, 2005. Fast webpage classification using URL features. Proceedings of the 14th ACM International Conference on Information and Knowledge Management, October 31-November 05, 2005, ACM, Bremen, Germany, ISBN: 1-59593-140-6, pp: 325-326.

Khonji, M., Y. Iraqi and A. Jones, 2011. Lexical URL analysis for discriminating phishing and legitimate websites. Proceedings of the 8th Annual Conference on Collaboration Electronic Messaging Anti-Abuse and Spam, September 01-02, 2011, ACM, Perth, Australia, ISBN: 978-1-4503-0788-8, pp: 109-115.

Khonji, M., Y. Iraqi and A. Jones, 2013. Phishing detection: A literature survey. IEEE Commun. Surveys Tutorials, 15: 2091-2121.

Ma, J., L.K. Saul, S. Savage and G.M. Voelker, 2011. Learning to detect malicious URLs. ACM Trans. Intellig. Syst. Technol., 2: 1-23.

Marchal, S., J. Francois, R. State and T. Engel, 2014. Phishstorm: Detecting phishing with streaming analytics. IEEE. Trans. Netw. Serv. Manage., 11: 458-471.

McGrath, D.K. and M. Gupta, 2008. Behind phishing: An examination of phisher modi operandi. Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats, April 15, 2008, San Francisco, CA, USA., pp: 1-8.

Thomas, K., C. Grier, J. Ma, V. Paxson and D. Song, 2011. Design and evaluation of a real-time url spam filtering service. Proceedings of the 2011 IEEE Symposium on Security and Privacy (SP), May 22-25, 2011, IEEE, California, USA., pp: 447-462.