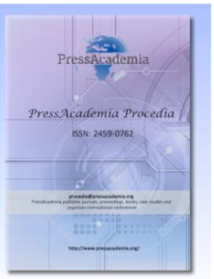
PressAcademia Procedia - (PAP), ISSN: 2459-0762

PressAcademia

PressAcademia

Procedia



2nd World Conference on Technology, Innovation and Entrepreneurship

May 12- 14, 2017, Istanbul, Turkey. Edited by Sefer Şener

DETECTING PHISHING WEBSITES USING SUPPORT VECTOR MACHINE ALGORITHM

DOI: 10.17261/Pressacademia.2017.582

PAP-WCTIE-V.5-2017(20)-p.139-142

Dogukan Aksu[1], Abdullah Abdulwakil[2], M. Ali Aydın[3]

[1]Istanbul Kültür University, Engineering Faculty, Computer Engineering Department, 34156, Bakırköy, Istanbul, Turkey. d.aksu@iku.edu.tr

[2]İstanbul University, Engineering Faculty, Computer Engineering Department, 34320 Avcılar, Istanbul, Turkey. abdullah19.right@gmail.com

[3]İstanbul University, Engineering Faculty, Computer Engineering Department, 34320 Avcılar, Istanbul, Turkey. aydinali@istanbul.edu.tr

ABSTRACT

Cybersecurity is one of the most important areas which aims to protect computers or computer systems, networks, programs and data from an attack such as; financial systems, biometric security systems, military systems, personal information security etc. Nowadays, there are a lot of rule-based phishing detection systems which are created to help people who can't understand which URL is real and which one is fake URL address. This paper proposes a method with supervised machine learning that classifies the URLs to legitimate and phishing. By using support vector machine (SVM) classification, a machine-learning algorithm, with an MATLAB-based computer program to give a warning message to the users about the reliability of the web page. In this paper, phishing detection system is implemented with SVM to avoid the internet users from becoming a victim of phishers to do not lose financial and personal information.

Keywords: Cyber security, phishing, machine learning, support vector machine, matlab

1. INTRODUCTION

In this fast growing modern technology driven world, the internet is one of the most important technology not only for individual users but also for organizations that have an online business. Most the organizations have an online business such as sales of product and services (Liu & Ye, 2001). This may put internet users to different types of risks which may result in loss of private information, identity theft, and financial losses (Abdelhamid, Ayesh, & Thabtah, 2014). Phishing is a method to deceive end-users to visit fraudulent web pages which have an almost same look and feel of the trusted web pages with the aim to steal personal and financial information. The information gained from the user such as usernames, passwords, social security numbers, and bank account details will give the attacker the ability, to imitate the victim and make financial transactions and put the victim in financial and emotional losses.

According to Anti-Phishing Working Group (APWG) report, In 2016 the total number of phishing attacks were 1,220,523 which show 65% increase over 2015 (Anti-Phishing Working Group, 2017). These websites can trap a lot of users and mislead them to expose their sensitive information, mainly the users who have less awareness about the safe usage of internet and do not inspect the web pages before exposing their personal information.

In this paper, we focused on web pages which have phishing URL address. Phishing is one of the major security issues in the modern technological world where all the information is in electronic format. To illustrate, if the user may not understand a phishing web page, the user may victimized of fraudulent act of phishing and phisher can have access to the bank account of the user and can withdraw money or the best case scenario is, if the phisher could not withdraw money at least he or she will have access to some valuable information of the victim.

Remaining parts of the paper is organized into the following sections. Section 2 explains a brief survey of the literature review. In Section 3, we present the features, SVM machine learning algorithm, and methodology. Section 4 highlights the finding and discussion. Finally, we conclude our paper in Section 5.

2. LITERATURE REVIEW

Numerous researchers have researched to eliminate phishing and guard online users against deceiving, but very few of these researchers are focusing on web page phishing detection. This section presents an outline of phishing detection methods. Broadly phishing detection is categorized into two, the first one is user education based and the second one is software based. Furthermore, the software based phishing detection is categorized into blacklist (heuristic), machine learning, and visual similarity. Machine learning methods utilize the features that are both common to phishing web pages and legitimate web pages such as length of the URL, Number of dots in the URL, At "@" symbol in URL, etc. In heuristic approach, the web page is considered as phishing if it matches the rules defined by the heuristic (Xiang, Hong, Rose, & Cranor, 2011). The drawbacks of phishing detection based on visual similarity and blacklist are, that they usually do not detect new phishing attack web pages(zero hour attack) (Jain & Gupta, 2016).

(Zhang, Hong, & Cranor, 2007) introduced CANTINA which is content-based approach and to reduce false positive rate some heuristics are also introduced.

3. DATA AND METHODOLOGY

Firstly, we create a dataset which consists of Phishing URLs have been taken from PhishTank (Phishtank) and for Non-Phishing URLs, the browsing history has been used.

Secondly, we defined the most known features in literature that are extracted from the dataset. These features are described as follows.

1- Long URL: Phishers are using long URLs to hide mistrustful parts in the address bar.

2- Dots: A legitimate URL can contain at most 5 dots. If a web page contains more than 5 dots it may be considered as phishing URL. For example, the following URL is considered as phishing because it contents more than five dots http://forccis.com.br/audit.com.verification.filling.information.ub/w2-form/ (Phishtank)

3- IP Address: IP addresses can be used instead of Fully Qualified Domain Name (FQDN). For example http://81.31.25.201/ok/rmf/pessoafisica.php (Phishtank). For security reasons legitimate websites do not use this method so if an IP Address exists in a web page link, it is declared as phishing. This features can be demonstrated as an important phishing detection feature because the online availability of phishing websites are few days.

4- SSL Connection: In general, web pages that have SSL Connection(HTTPS) can be considered non-phishing.

5- At (@) Symbol: Phishers may use @ symbol to give the impression of a legitimate internet address. Using @ symbol in the URL ignores whatever is there before @. This allows the phishers to write a legitimate URL before @ and hide the fraudulent part of the URL, for example, https://mail.google.com/mail/@http://www.phish1.com. Therefore if URL contains @ symbol it is considered as phishing.

6- Dash (-) symbol: The dash symbol is not commonly used in legitimate URLs. Phishers are adding suffix or prefix separated by dash (-) symbol to misguide the user about the originality of the web page, for example, http://appleid-icloud.in (Phishtank) is considered as phishing web page.

Innumerable real-world learning tasks have solved successfully by using support vector machine (SVM) that is machine learning algorithm developed by Vapnik (Cortes & Vapnik, 1995). In this study, binary classification has been used because we have had two different types of URLs phishing or non-phishing. Proposed algorithm in this paper is presented in **Figure1**.

Figure 1: The Design of Phishing Detection System

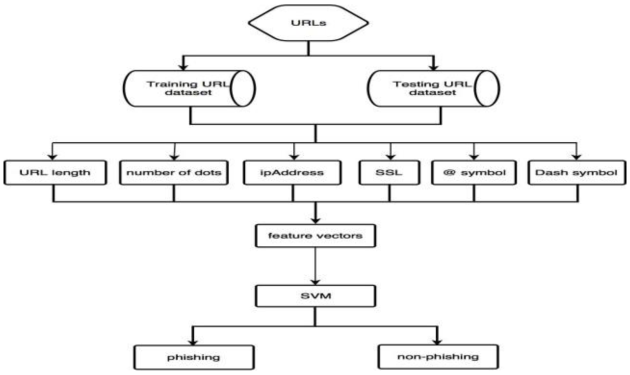


Figure1 represents the design of our system which is demonstrated as follows.

1) Phishing URLs have been taken from PhishTank (Phishtank) and for Non-Phishing URLs, the browsing history has been used.
2) In feature selection, six features which are URL length, the number of dots, ipAddress, SSL connection, at symbol(@), dash symbol(-) have been selected.
3) The selected features extracted from the dataset.
4) In training, SVM is trained via above-extracted features.
5) URLs which are not included in the training dataset are given to SVM classifier for testing.
6) SVM classifier returns either an URL is phishing or non-phishing. The confusion matrix for phishing detection can be seen in Table 1 (Fang, Koceja, Zhan, Dozier, & Dipankar, 2012)

Table 1: The Confusion Matrix for Phishing Detection

	DP	DNP
AP	TP	FN
ANP	FP	TN

- AP: Actual Phishing
- ANP: Actual Non-Phishing
- DP: Detected as Phishing
- DNP: Detected as Non-Phishing

- TP: True Positive
- TN: True Negative
- FP: False Positive
- FN: False Negative

4. FINDINGS AND DISCUSSIONS

The algorithm presented in the study is implemented using MATLAB programming language. The output of the system is given in Table 2.

Table 2: Output of the System

	DP	DNP
AP	22	3
ANP	2	73

Results in Table2 shows that the percentage of detected as phishing to the actual phishing is 88 % while the percentage of detected as non-phishing to the actual phishing is 12 %. Similarly, the percentage of detected as phishing to the actual non-phishing is 0.02 % while the percentage of detected as non-phishing to the actual non-phishing is 97 %.

Table 3: Performance Measurement

Measure	**Formula**	**Result(%)**
Accuracy	(TP+TN) / (TP+FP+FN+TN)	95
Recall	TP/(TP+FN)	88
Precision	TP/(TP+FP)	91,66
F1 score	2TP/(2TP+FP+FN)	89,79

Accuracy, recall, precision and f1 score are used for performance measurement (Shouval, Bondi, Mishan, Shimoni, Unger, & Nagler, 2014). The performance measurement of our system is shown in **Table3.** As it can be seen that accuracy, recall, precision and f1 score rates are 95 %, 88 %, 91.66 % and 89.79 % respectively. In this paper, by decreasing the number of features to 6 we have been able to achieve approximate accuracy to the study of (Akanbi, Amiri, & Fezaldehkordi, 2015)which has been obtained using 9 features. Despite the fact that we have decreased the number of features, our system produces promising results.

5. CONCLUSION

In this paper, the method that provides a detection system to prevent the users from being a phishing victim is presented. The first dataset has been created using Phishtank for phishing URLs and non-phishing URLs have been taken from browser history. Specified features are extracted from the training dataset and used for classification of URLs to phishing and non-phishing. By using SVM classification algorithm, phishing URLs are detected. This method can detect phishing URLs but sometimes when the URL contains a feature which is not available in our features, the system gives the wrong result. Furthermore, features can be changed and affect the result according to URLs, therefore features must be defined dynamically. In future work, dynamically selecting the best features can be handled using a deep learning algorithm to get more accurate results.

REFERENCES

Abdelhamid, N., Ayesh, A., & Thabtah, F. (2014). Phishing detection based Associative Classification data mining. *Expert Systems with Applications*, 5948-5959.

Akanbi, O. A., Amiri, I. S., & Fezaldehkordi, E. (2015). *A Machine Learning Approach to Phishing Detection and Defense.* ELSEVIER.

Anti-Phishing Working Group, J. (2017, Feb. 23). *Phishing Activity Trends Report, 4th Quarter 2016.* Retrieved March 10, 2017, from APWG: https://docs.apwg.org/reports/apwg_trends_report_q4_2016.pdf

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning. 20(3): 273-297.

Fang, X., Koceja, N., Zhan, J., Dozier, G., & Dipankar, D. (2012). An Artificial Immune System for Phishing Detection. *IEEE World Congress on Computational Intelligence*.

Jain, A. K., & Gupta, B. B. (2016). Comparative Analysis of Features Based Machine Learning Approaches for Phishing Detection. *International Conference on Computing for Sustainable Global Development (INDIACom)*, (pp. 2125-2130).

Liu, J., & Ye, Y. (2001). Introduction to e-commerce agents: marketplace solutions, security issues, and supply and demand. *In E-commerce agents, marketplace solutions, security issues, and supply and demand*, 1-6.

Phishtank. (n.d.). Retrieved February 9, 2017, from OpenDNS: http://www.phishtank.com

Shouval, R., Bondi, O., Mishan, H., Shimoni, A., Unger, R., & Nagler, A. (2014). Application of machine learning algorithms for clinical predictive modeling: a data-mining approach. *Bone Marrow Transplantation*, 49, 332–337.

Xiang, G., Hong, J., Rose, C. P., & Cranor, L. (2011). Cantina+: A feature rich machine learning framework for detecting phishing web sites. *ACM Transactions on Information and System Security (TTSSEC)*, 14(2): p. 21.

Zhang, Y., Hong, J., & Cranor, L. (2007). Cantina: a content-based approach to detecting phishing web sites. *Proceedings of the 16th international conference on World Wide Web.*