Mobile Phishing Websites Detection and Prevention Using Data Mining Techniques

https://doi.org/10.3991/ijim.v13i10.10797

Huda Yousif (✉), Karim Hashim Al-saedi, Mustafa Dhiaa Al-Hassani
Mustansiriyah University, Baghdad, Iraq
hy94@uomustansiriyah.edu.iq

Abstract—The widespread use of smart phones nowadays makes them vulnerable to phishing. Phishing is the process of trying to steal user information over the Internet by claiming they are a trusted entity and thus access and steal the victim's data (user name, password and credit card details). Consequently, the need for mobile phishing detection system has become an urgent need. And this is what we are attempting to introduce in this paper, where we introduce a system to detect phishing websites on Android phones. That predicts and prevents phishing websites from deceiving users, utilizing data mining techniques to predict whether a website is phishing or not, relying on a set of factors (URL based features, HTML based features and Domain based features). The results show system effectiveness in predicting phishing websites with 97% as prediction accuracy.

Keywords—Mobile Phishing attacks; Phishing; Data Mining; Web-based Phishing attack.

1 Introduction

Nowadays the Internet has become another world in itself, providing all necessities with a click of a button. The presence of on-line services like e-banking, e-marketing, etc., has made the people life more convenient by allowing them to manage their transactions sitting at home. In return exposed them to enormous security threats. Presently it becomes even more easier to acquire these services, due to (the vast spread of smart-phones and tablets, its ease of use in addition to its portability and their ability to provide services that PCs provides). According to statistics in October 2018, the number of internet users worldwide has reached 2 billion. With more than 1 billion people using mobile phones [1]. There are many types of security menace on the Internet such as: Phishing, malware attack, man-in-the-middle attack, etc., and Phishing is the most predominant of all [2]. Phishing can be defined as a deceptive act that used to deceive users over the internet with the aim of acquiring their personal information. According to Kaspersky Lab, since the first quarter of 2015 an increase of around a million phishing cases have been reported [3]. Recently, dealing with websites has increased significantly as it provides all the services that users are looking for, from online banking, shopping, socializing and much more. Which made

the Internet users at the present time familiar with being requested to provide their information to these websites, that why most of phishing attacks are in the form of a fraudulent website [4]. As we have mentioned that phishing attacks have been appeared since many years ago, the first of which appeared in 1995 when the first time the word phishing was coined [5]. Since then, many researchers and companies have tried to address these attacks to detect, prevent and educate these phishing attacks. For the detection purpose, there are certain methods have been applied such as: blacklist method, heuristic techniques and visual appearance methods. In the blacklist method, the website is checked in a list of blacklisted websites, whereas this method may be effective in the fast detection of blacklisted websites. But it has a significant drawback as it cannot detect the websites that appears merely for a day or a couple of days or even a few hours "zero-day phishing attack"[6]. As for the heuristic techniques, it is considered more effective in compare with the blacklist in dealing with zero-day phishing attacks. Since it depends on the features extracted from (URL, HTML, and Search Engine), as well as the data mining technique that used to determine the status of the website [7]. Whereas visual appearance method depends on the similarity between websites in phishing detection. When the suspected websites match with the legitimate website visual characteristics, it checks if the URL is in the authentic domain URLs list. If it's not found in the list then it is marked as a phishing website [7]. In this paper we try providing a system that uses the least possible number of features and the most effectives ones in predicting "zero-day" phishing websites on smart phones, and preventing them from deceiving users.

2 Mobile Phishing Websites

Mobile devices facilitate phishing attacks due to the following properties: Firstly: the rapid increase of mobile users worldwide. This attracted phisher to shift their techniques to mobile devices. Researchers from Trend Micro in 2012 found 4000 phishing URLs for mobile web pages [8]. Secondly: the limited screen sizes makes it difficult for mobile users to determine legitimate web-page from phishing one. It also makes browsers capable of hiding the complete URL of the requested page, hence helping the phisher deceiving the mobile user. Due to the mobility nature of the mobile phones, users tend to respond to interactions with less concentration which might yield to approving a phishing process. Phishers and Malware creators are aware of these attractive properties and hence have moved their efforts to mobile devices. Mobile websites considered as the ideal environment for phishing attacks, for the following reasons:

- Web pages are easily forged by copying the source code of the site for falsification
- Complex anti-phishing techniques cannot run efficiently on mobile phones because of mobile devices limited resources
- Mobile phones have small screen sizes; therefore, it is difficult to detect phishing websites depending on their appearance or via security indicators
- Browsers on mobile phones are lightweight and their security capabilities have been reduced, to suit mobile phones abilities

This what makes the focus of this research on website phishing attacks detection and prevention.

3 Related Work

There are a large number of literature available in phishing detection methodologies for computer devices (See [6][7] for a comprehensive survey). Concerning phishing detection on mobile devices: Authors in [9] presents the phishing attacks on android and IOS web browsers and an assessment of the available protection techniques against them, they compare the protection they offer with their desktop counterparts and analyze the gap between the two. Hossain, Tulin Klintic and Victor in [10], propose a phishing detection taxonomy for mobile environment, endeavoring to delineate and discuss every single conceivable situation of phishing assaults and related countermeasures too. Leaving aside all possible attacks related to very specific vectors (e.g. Bluetooth, Smishing and Vishing), the paper emphasizes the lack of solutions dedicated to mobile devices other than black/white lists. A risk assessment on mobile platforms has been proposed in [11], where a study conducted on 85 sites and 100 applications found that sites and applications consistently ask users to type their passwords into context of being susceptible to impersonation. The usage of test phishing assaults on the Android and IOS stages exhibited that attackers can parody legitimate applications precisely, suggesting that the danger of phishing assaults on mobile phones is more noteworthy than it has beforehand been valued. In [4] the authors presented phishing scheme for mobile phones, attempting to utilize OCR content extraction tools, to confirm the authenticity of a website and compare the content that extracted from a login form with the corresponding second-level domain name (SLD). This stems from the presumption that most known enterprises utilize their brand name as the SLD of their official sites, which is likewise utilized as a picture inside their login forms. However, the introduced framework also uses white lists and heuristics to diminish false positives or increase the effectiveness of the overall framework (e.g. authentic websites always utilize domain names as verification of their identities while phishers are predominantly use their IP in URL to conceal their fake identities).

4 Proposed System

The proposed system is called "**Phishield**" which is a merge of "Phish" and "Shield". Phishield (Figure 1) is divided into four stages. The first stage is the **system Database (SDB)** checking, where a system database built by the detection system, to check if that website is already checked, to minimize time and resource consuming. And the second stage is checking if the Domain name is an IP. Stage three feature extraction. And stage four that classify and predict whether a website is phishing or not and finally take the proper action towards the website.

4.1 SDB checking

To minimize time wasting and system resources consuming, a system Data Base SDB have been utilized. As mobile user tries to browse for certain website, the system start functioning when user click on link and wait the browser loading the website, meanwhile the system will fetch the URL and start its analyzing process. System first will check if that URL exist in SDB and take the proper action at the same time. SDB would contain each Website URL have been Checked by the Detection System and its type (i.e. Phishing or Legitimate) so the system can take the proper action without having to resort to the steps that follows. Which contribute in the fast detection of phishing URLs.

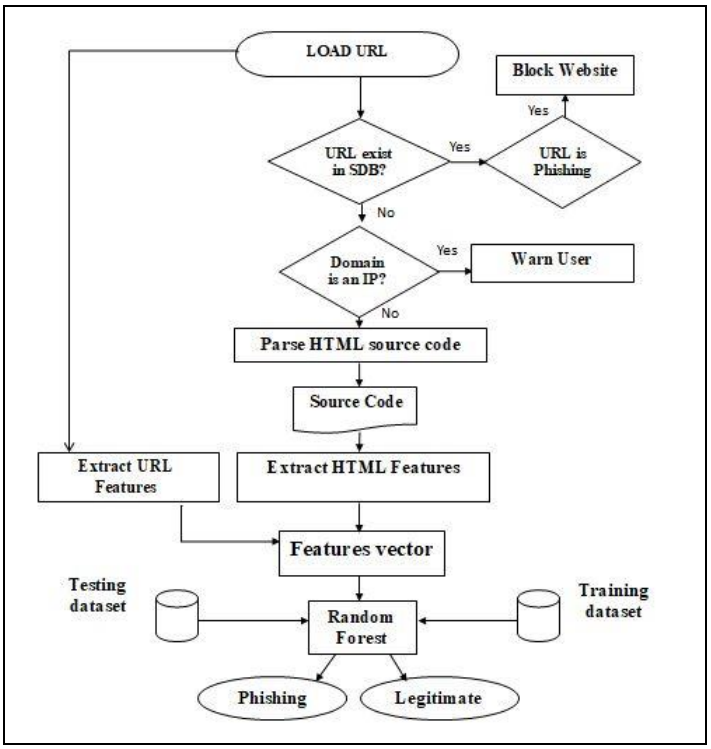

Fig. 1. Proposed System

4.2 Check if the domain name is an IP

As a verification of their identities, legitimate websites use their company, institute or services names as a domain name. While phishers are inclined to use their IP addresses as an alternative of a domain name [4] (e.g. http://209.97.129.236/). Thus, system before loading the website will analyze the link and extract the domain name of the website (Figure 2, shows the structure of a URL). And check if an IP address is used as a domain name of a website, if it is so then system will alert user that site

might be phishing website and give him the choice to stop the loading process of the website or keep browsing, because it's not necessarily be a deceptive website. If it is not an IP, System moves to the next phase.

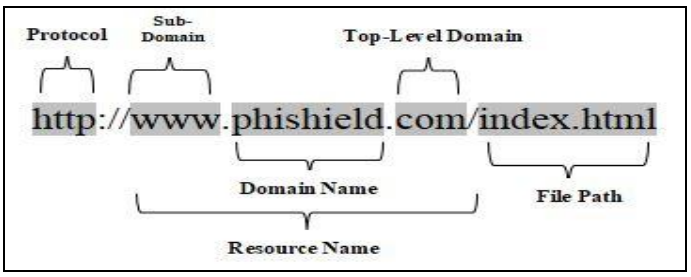

Fig. 2. URL Structure

4.3 Features extraction

In this phase system will collect set of features that will based on, in making decision about the website. The features used in the proposed System are selected by studying the most common 30 attributes in the area of phishing websites detection, then by selecting the most effective 14 features, based on information gain algorithm. These features are collected from the following publications: (M.Rami [12], M.Khonji [6], S.Lakshim [13]). The selected set of features demonstrated their effectiveness in the right diagnosing of the website status. And they are divided into three subgroups, Table 1 present the groups of features.

Table 1. Phishing Websites Features used in Phishield

Features Groups	No.	Feature
URL-based Features	1	Long URL
	2	Using @ Symbol
	3	Prefix or Suffix Separated by "-"
	4	Sub-Domain and Multi Sub-Domains
	5	HTTPS and SSL "Secure Sockets Layer"
	6	Slashes in page address and URL
	7	URL is Blacklisted
HTML-based Features	8	Request URL
	9	URL of Anchor
	10	Server Form Handler (SFH)
	11	Redirect Page
	12	Nil Anchor
Domain-based Features	13	Age of Domain
	14	Website Traffic

5 Analysis and Assessment

System has been applied on **496** instance of training dataset (collected from Phishtank [14] and Alexa [15]). Table 2 shows a comparison of the trained model's performance in predicting test-set classes, that evaluated based on two criteria, the prediction accuracy and the training time. The classification algorithms, Random Forest, Decision Table, C4.5 (J48), Support Vector Machine (SMO) and Bayes Net are implemented and trained using WEKA (open source java application that implements a collection of machine learning algorithms and data pre-processing tools [16]).

Table 2. Performance comparison of classifiers

Classifiers	Time taken to build model	Correctly predicted instance	Incorrectly predicted instance	Prediction accuracy
Random Forest	0.19s	485	11	97.7823 %
Decision Table	0.15s	469	27	94.5565 %
J48	0.23s	466	30	93.9516 %
SMO	0.17s	465	31	93.75%
Bayes Net	0.2s	452	44	91.129 %

The results showed that Random Forest algorithm outperformed Decision Table, C4.5(J48), Support Vector Machine (SMO), Bayes Net in predicting Phishing and Legitimate Websites. Not the fastest but it gaves the most accurate predictions in a nearby interim of time taken by other algorithms to build the model.

6 Performance Evaluation

We use **Android Studio 3.4** as a tool for building the proposed system. The application has been implemented as a web browser application on Samsung and nexus smartphones running **Android 5.1(Lollipop)** operating system. In our experiments Phishield has been tested in online real-time mode, applying the application on a real legitimate website and phishing website (logged into from Phishtank), Phisheild effectively marked legitimate websites and allow to receive the browsing process with no problems, and at the same time detected all phishing websites as a phishing attack and prevent us "as user" from loading the websites. Figure 3 and Figure 4 shows different cases handled by Phishield.

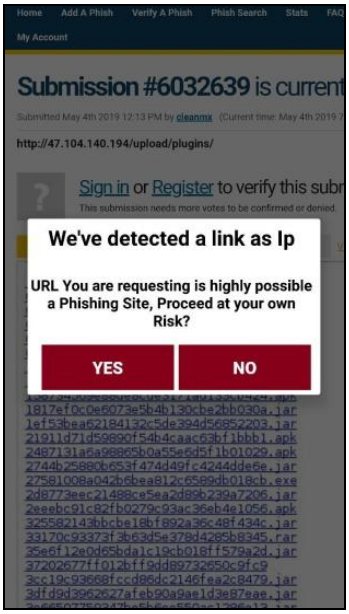

Fig. 3. Phishield Alert of a website that uses an IP address

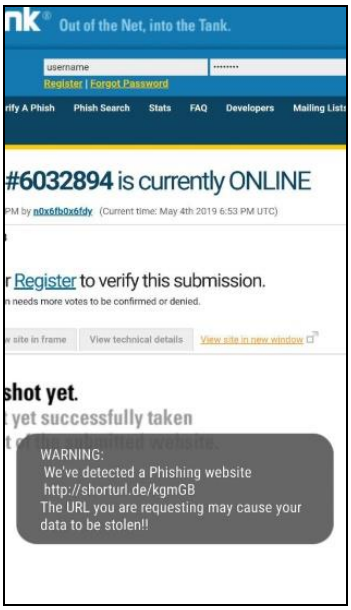

Fig. 4. Phishiled Detected a Phishing website

7 Conclusion

This work models the prediction of phishing websites on mobile devices as a classification task and demonstrate the machine learning approach to predict the websites status and take the proper action towards it. Random Forest, Decision Table, C4.5 (J48), SMO and Bayes Net classifiers have been applied on the training dataset and their performance has been evaluated based on their results in predicting the websites status. Evaluation results have shown Random Forest classifier outperforms other models and produces best results. Hopefully, the detection system would be applied onto mobile common web browsers "Chrome", "Android-built in browser" and other browsers used by users in the future also applying more strict actions towards phishing websites.

8 References

[1] We Are Social. [Online].Available from: https://wearesocial.com/au/special-reports/the-state-of-the-internet-in-q4-2018.

[2] Joshi, Rushikesh, "Interactive Phishing Filter" (2015). *Master's Projects*. 430. https://scholarworks.sjsu.edu/etd_projects/430.

[3] Kaspersky Lap. [Online].Available: https://www.kaspersky.com/blog/spam-and-phishing-in-q1-2015-banks-and-banking-trojans/15075/.

[4] Longfei Wu, Xiaojiang Du, and Jie Wu, "MobiFish: A Lightweight Anti-Phishing Scheme for Mobile Phones", in 23rd International Conference on Computer Communication and Networks (ICCCN), 4-7 Aug. 2014, Shanghai.

[5] Rekouche, Koceilah. "Early Phishing." *CoRR* abs/1106.4692 (2011).

[6] Mahmoud Khonji, Youssef Iraqi, Andrew Jones, "Phishing Detection: A Literature Survey", in IEEE COMMUNICATIONS SURVEYS & TUTORIALS, Vol. 15, No. 4, Fourth Quarter 2013, pp. 2091-2121. https://doi.org/10.1109/SURV.2013.032213.00009

[7] Gaurav Varshney, Manoj Misra, Pradeep K. Atrey, "A survey and classification of web phishing detection schemes", in SECURITY AND COMMUNICATION NETWORKS, Vol. 9, 10.1002/sec.1674, October 2016. https://doi.org/10.1002/sec.1674

[8] Trend Micro, "Mobile phishing: A problem on the horizon," (2012).

[9] Nikos Virvilis, Nikolaos Tsalis, Alexios Mylonas, Dimitris Gritzalis, "Mobile devices: A phisher's paradise", in 11th International Conference on Security and Cryptography (SECRYPT 2014), 28-30 Aug. 2014, Vienna, Austria. https://doi.org/10.5220/0005045000790087

[10] Hossain Shahriar, Tulin Klintic, Victor Clincy, "Mobile Phishing Attacks and Mitigation Techniques", in Journal of Information Security, 2015, 6, pp 206-212, Publishe Online July 2015 in Scientific Research Publishing,. https://doi.org/10.4236/jis.2015.63021

[11] Adrienne Porter Felt, David Wagner, "Phishing on Mobile Devices", in Web 2.0 Security and Privacy Workshop (W2SP), May 26 2011, Oakland, CA.

[12] Mohammad, Rami, McCluskey, T.L. and Thabtah, Fadi Abdeljaber (2014) Intelligent Rule based Phishing Websites Classification. IET Information Security, 8 (3). pp. 153160. ISSN 17518709. https://doi.org/10.1049/iet-ifs.2013.0202

[13] Santhana Lakshim, Vijaya MS, "Efficient Prediction of Phishing Websites using Supervised Learning Algorithms", in International Conference On Communication

Technology and System Design 2011, 10.1016/j.proeng.2012.01.930. https://doi.org/10.1016/j.proeng.2012.01.930

[14] Phish Tank. [Online].; Available from: https://www.phishtank.com/.

[15] Alexa the web Information Company. [Online]. [Cited 2012 January 26. Available from: https://www.alexa.com/.

[16] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. Waikato Environment for Knowledge Analysis. [Online]. Available from: https://www.cs.waikato.ac.nz/ml/weka/

9 Authors

Huda Yousif did Bachelor of computer science from University of AL-Mustansiriyah at 2016. Now I'm a Master degree student in AL-Mustansiriyah University, in Data Mining Techniques and Internet security.

Dr. Karim Hashim Al-Saedi, received the PhD degree in computer science (Data Mining and Network Security) from the University Sains Malaysia (USM) in 2013, and M.Sc. in computer science at University of Technology, Baghdad, Iraq, in 2005, He is working as Lecture in Department of Computer Science, College of science, at AL-Mustansiriyah University, Baghdad, Iraq. His research interests include the areas of Data Mining, Advanced Internet Security and monitoring, Medical Image Assessment, and Machine learning. He published 21 papers. He is a Member of the Internet Society (ISOC), since 2012. Malaysia, Member of the UNCTAD Virtual Institute, since 2006. Geneva, Switzerland, and Member of the Iraq Computer Society since 1997. Baghdad, Iraq.

Dr. Mustafa Dhiaa Al-Hassani received the Ph.D. degree in computer science (Identification Techniques using Speech Signals and Fingerprints) from Al-Nahrain University in 2006 (1st Rank), and MSc. in computer science (Design of a Fingerprint Recognition System using Wavelet Transformation) from Al-Nahrain University in 2002. He is working as Lecture in Computer Dept./ College of Sciences/ Mustansiriyah University, since 2007. Get many International Certificates in the field of IT. Occupy several administrative positions. Publish a lot of Papers and Books in computer science fields. His interest in the area: Information Security, Multimedia Processing, Data Compression, Databases, Pattern Recognition, eLearning.

Article submitted 2019-06-01. Resubmitted 2019-07-15. Final acceptance 2019-07-16. Final version published as submitted by the authors.