DETECTION OF PHISHING WEBSITES USING MACHINE LEARNING

SHWETHA¹ & PROF. KAVITHA S.N²

¹Student, Department of Information Science and Engineering, RV College of Engineering, Bengaluru, India

²Assistant Professor, Department of Information Science and Engineering, RV College of Engineering, Bengaluru, India

ABSTRACT

Trying to access personal information nowadays has become more common. Phishing is an attack where the hackers take advantage of the trust factor of the target and try to gather sensitive information of a target such as a username, password, etc. by disguising as a trustworthy entity. There are many anti-phishing methods such as blacklist, heuristic, visual similarity and, machine learning. The blacklist method is widely used because it is easy to use and execute, but it fails to detect new phishing attacks. This paper proposes a methodology of phishing identification framework where various machine learning algorithms like random forest, support vector machine, logistic regression are used for the comparison conciseness to predict more accuracy. It also includes data analysis, data visualization, and, detecting the phishing website. After detailed research, we proposed a framework that overcomes the disadvantages of other approaches.

KEYWORDS: Phishing Detection, Feature Extraction, Phishing Website, Phishing Attacks

Received: Jun 09, 2020; **Accepted:** Jun 29, 2020; **Published:** Aug 06, 2020; **Paper Id.:** IJMPERDJUN2020644

INTRODUCTION

Phishing is a social manipulation assault aimed at leveraging the vulnerability found in the program at the end of the user. For example, a program may be technically secure enough for password theft, but an unrecognized user can leak his / her password when an attacker sends a request for a false password update via a fake website. To resolve this problem, a layer of security must be added for use.

As of late, there have been a few examinations that attempted to tackle the phishing issue. A few analysts utilized the URL furthermore, contrasted it and, existing boycotts that contain arrangements of vindictive sites, which they have been making, and others have utilized the URL in a contrary way, to be specifically contrasting the URL and a whitelist of real sites.

The latter approach uses heuristics, which is used Database of signatures for any known attacks that match the Signature of the heuristic template to determine whether it's a phishing This is the platform. Also besides tracking traffic on Alexa 's website is another way in which researchers have been applied to detect websites for phishing.

In this article, the emphasis will be on the combination of features So we get the Random Forest (RF) strategy, because it does. High precision, fairly stable and, good results.

LITERATURE SURVEY

Blacklist Approach and Whitelist Approach

In [1], Pawan Prakash, Manish Kumar, Ramana Rao Kompella, Minaxi Gupta (2010) proposed a prescient boycott way to deal with recognize phishing sites. It distinguished a new phishing URL utilizing heuristics and by utilizing a proper coordinating calculation. Heuristics made new URL's by joining portions of the known phished sites from the accessible boycott. The coordinating calculation at that point figures the score of URL. If this score is over a given edge esteem it hails this site as a phishing site. The score was assessed by coordinating different pieces of the URL against the URL accessible in the boycott. Maintaining the Integrity of the Specifications.

In [14], Jung Min Kang and DoHoon Lee described approach which detected phishing based on users online activities. This method maintained a white list as a part of the users' profile. This profile was dynamically updated whenever a user visited any website. An engine used here identified a website by evaluating a score and then comparing it with a threshold score. The score was calculated from the entries available in the user profile and details of the current website.

Heuristic Approach

In [7], Aaron Blum, Brad Wardman, Thamar Solorio proposed a work that focused on the exploration of surface-level features from URLs to train a confidence weighted learning algorithm. The idea is to restrict the source of possible features to the character string of the URL and avoid having the vulnerability of extracting host-based information. Every URL is displayed as a vector of binary features. These vectors are fed to the online algorithm were at the time of testing, previously unseen URLs in the binary feature vector is then mapped to it. The learner continues this new vector and output into the final result, either phish or legitimate.

In [15], Guang Xiang, Jason Hong, Carolyn P. Rose, Lorrie Cranor proposed CANTINA+, a comprehensive feature-based approach in the literature including eight novel features, which exploits the HTML Document Object Model (DOM), search engines and third party services with machine learning techniques to detect phish. Also, two other filters are designed in it to help reduce FP and achieve good runtime speedup. The first is a near-duplicate phish detector that uses hashing to catch highly similar phish. The second is a login form filter, which directly classifies webpages with no identified login form as legitimate.

In [8], Joby James, Sandhya L, Ciza Thomas proposed a work which with the combined help of the blacklisting approach and the Host-based Analysis applied certain classifiers that can be used to help detect and takedown various phishing sites. The host-based, popularity based and lexical based feature extractions are applied to form a database of feature values. The database is knowledge mined using different machine learning methods. After evaluating the classifiers, a particular classifier was selected and was implemented in MATLAB.

In [9], APWGM published a case study citing the importance of the WHOis tool and how invaluable it has been for the rapid phishing site shutdown over the past few years all around the globe.

Visual Similarity Approach

In [2], A. Mishra and B. B. Gupta presented a hybrid solution based on URL and CSS matching. In this approach, it can detect embedded noise contents like an image on a web page which is used to sustain the visual similarity on the webpage.

They used the technique used in [3] by Jian Mao, Pei Li, Kun Li, Tao Wei, and Zhenkai Liang to compare the CSS similarity and used it in their technique. The different types of visual features are - text content and text features. Text features are like font color, font size, background color, font family, and so forth. This approach matches the visual features of different websites because the attacker copies the page content from the actual website.

In [5] Matthew Dunlop, Stephen Groat, and David Shelly proposed a browser-based plug in called goldfish to identify phishing websites. It uses the website logos to identify the fake website. The attacker can use the real logo of the target website to trap the internet users. Three stages to it are:

- Logo Extraction: Goldfish is used to extract the website logo from the suspicious website. Then it converts it into text using optical character recognition (OCR) software.
- Legitimate website extraction: The text obtained is used as a query for the search engine. Generally, search engine “google” is used because it always return genuine websites in their top results.
- Comparisons: Suspicious website is compared with the top result obtained from the search engine based on different features. If any domain is matched with the current website then it is declared legitimate or else make it phishing site.

PROPOSED WORK

The suggested framework consists of pre-processing, data interpretation, data visualization, and the identification of whether the URL is a phishing website or a legitimate one. Used three machine learning algorithms (LR), Random Forest (RF), vector support (SVM) to identify websites as legitimate and phishing.

Preparing the Dataset

The data set is given to the machine learning model based on the data set of the model being educated. Each new detail-filled at the time of the application form serves as a test data collection. After research is carried out, the model prediction based on the inference concludes based on the training data sets. Studying the layout of a phishing URL and a real URL is used to identify a phishing URL.

Table 1

Variable	Description
having_IP_Address	Domain IP
URL_Length	URL length
Shortening_Service	Tiny URL
having_At_Symbol	@ symbol
double_slash_redirecting	Check for //
Prefix_Suffix	-
having_Sub_Domain	Having dot
SSLfinal_State	connection state
Domain_registration_length	Domain length
Favicon	-
port	Port no.
HTTPS_token	Having HTTPS
Request_URL	-
URL_of_Anchor	-
Links_in_tags	-
SFH	-

Submitting_to_email	-
Redirect	-
on_mouseover	-
RightClick	-
popUpWindow	--
Iframe	-
age_of_domain	-
DNSRecord	-

Analysis Phase

Here three algorithms are used for the analysis part that is logistic regression, random forest, and support vector machine. The comparison is made among them to predict more accuracy and choose the best algorithm.

For different apps, we lay down various rules based on the study of phished and non-phished websites scraped off the internet.

Fig.1 and Fig 2, indicates the number of hyphens on the phished and legal websites. The Y-axis denotes the number of websites and the X-axis denotes the number of hyphens on the page. Based on this analysis, we concluded that phished websites do consist of hyphens in the domain part of the URL and that legitimate websites do not.

Figure 1: Hyphen Count of Phished Websites.

Figure 2: Hyphen Count of Legitimate Websites.

Data Visualisation

To visualize the given dataset in the form of graphical representation like pair plot, heat map, bar chart, pie-chart from matplotlib, seaborn library packages.

Figure 3: Pie Chart.

Requirement Analysis

- Software Requirements
 - Python 3.6
 - Anaconda Navigator
 - Scikit-learn (Package in Python)
 - Browser (Chrome)
- Hardware Requirements
 - Windows 7 above
 - Hard disk of at least 64 GB

System Architecture

Figure 4: System Architecture.

Design Phase

The flow of the proposed system is shown in Fig. 5

Figure 5: Flow Chart of the Proposed System.

RESULTS

The linear regression plot of expected output versus predicted output is shown in Fig. 5. This was predicted by the random forest algorithm. It has a slight deviation from the expected output for the phished websites.

Figure 6: Linear Regression Plot of Original Output Versus Predicted Output.

Figure 7: Machine Learning Accuracy Bar Plot.

The true positive, false positive, true negative, false negative count and accuracy results of 9076 test websites is as shown in Table 2

Table 2: Confusion Matrix Results

Algorithm	TN	TP	FP	FN	Accuracy
Logistic Regression	6447	2287	325	17	96.23 %
Decision Tree	6393	2341	326	16	96.23 %
Random Forest	6392	2374	297	13	96.58 %

CONCLUSIONS

The proposed system framework empowers the web clients to have a protected perusing and safe exchanges. It causes clients to spare their significant private subtleties that ought not be spilled. Giving our proposed framework to clients as augmentation makes the procedure of delivering a framework a lot simpler. The outcomes focus on the proficiency that can be accomplished utilizing the mixture arrangement of heuristic highlights, visual highlights and, boycott and whitelist approach and, taking care of these highlights to AI calculations. A specific challenge in these spaces is that crooks are continually making new techniques to counter our barrier measures. To prevail in this unique situation, we need calculations that constantly adjust to new models and highlights of phishing URLs. What's more, accordingly we utilize internet learning algorithms. This new framework can be intended to benefit the most extreme exactness. Utilizing various methodologies through and through will upgrade the precision of the framework, giving an effective security framework. The downside of this framework is distinguishing of some negligible bogus positive and bogus negative outcomes. These downsides can be disposed of by acquainting a lot more extravagant component with feed to the AI

REFERENCES

1. Ankit Kumar Jain and B. B. Gupta, "Phishing Detection Analysis of Visual Similarity-Based Approaches", Hindawi 2017.
2. A. Mishra and B. B. Gupta, "Hybrid Solution to Detect and Filter Zero-day Phishing Attacks", ERICA 2014.

3. Jian Mao, Pei Li, Kun Li, Tao Wei, and Zhenkai Liang, "Bait Alarm Detecting Phishing Sites Using Similarity in Fundamental Visual Features", INCS 2013.
4. Eric Medvet, EnginKirda and Christopher Kruegel, "Visual Similarity-Based Phishing Detection", ACM 2015. [5] Matthew Dunlop, Stephen Groat, and David Shelly, "GoldPhish Using Images for Content-Based Phishing analysis", IEEE 2010.
5. Haijun Zhang, Gang Liu, Tommy W. S. Chow, and Wenyin Liu, "Textual and Visual Content-Based Anti-Phishing A Bayesian Approach", IEEE 2011
6. Aaron Blum, Brad Wardman, Thamar Solorio, Gary Warner; "Lexical Feature-Based Phishing URL Detection Using Online Learning", Department of Computer and Information Sciences The University of Alabama at Birmingham, Alabama, 2016
7. Pawan Prakash, Manish Kumar, Ramana Rao Kompella, MinaxiGupta, Purdue University, Indiana University "PhishNet: Predictive Blacklisting to Detect Phishing Attacks".
8. The Anti-Phishing Working Group, DNS Policy Committee;" Issues in Using DNS Whois Data for Phishing Site Take Down",The Anti Phishing Working Group Memorandum, 2011.
9. Guang Xiang, Jason Hong, Carolyn P. Rose, Lorrie Cranor,"CANTINA+: A Feature-rich Machine Learning Framework for Detecting Phishing Web Sites", School of Computer Science Carnegie Mellon University, ACM Society of computing Journal, 2015.
10. Joby James, Sandhya L, Ciza Thomas "Detection of phishing websites using Machine learning techniques", 2013 International Conference on Control Communication and Computing (ICCC).
11. Mohsen Sharifi and Seyed Hossein Siadati "A Phishing Sites Blacklist Generator".
12. JungMin Kang and DoHoon Lee "Advanced White List Approach for Preventing Access to Phishing Sites". [14] Y. Zhang, J. I. Hong, and L. F. Cranor. Cantina: a content-based approach to detecting phishing web sites. In WWW '07: Proceedings of the 16th international conference on World Wide Web, pages 639– 648, New York, NY, USA, 2007. ACM.
13. Bansode, Ulka M., and Gauri R. Rao. "Study of Various Anti-Phishing Approaches and Introducing an Improved Method for Detecting Phishing Websites." *International Journal of Computer Science and Engineering (IJCSE)* 2.4 (2013): 151-156.
14. Hemanth, V., M. Shareef, and K S Ranjith. "Anti-Phishing Using Visual Cryptography." *International Journal of Computer Science and Engineering (IJCSE)* 2.3 (2013):21-26.
15. Maidamwar, Priya, Nekita Chavhan, and Uma Yadav. "Internet of Things: A Review on Architecture, Security Threats and Countermeasures." *International Journal of Computer Networking, Wireless and Mobile Communications (IJCNWMC)* 8.1 (2018):1-10.
16. Shaout, Adnan, and Ryan Banksto. "Enterprise it Logging in the Cloud: Analysis and Approaches." *International Journal of Computer Science and Engineering (IJCSE)* 3.2 (2014):47-66