Phishing Websites Detection

M. Srinithya[1], V. Ragasree[2], K. Sri Harshini[3]

Abstract: ***There is a range of customers who buy goods on the internet and make orders through numerous websites. There are many sites that require clients to provide touchy content on a regular basis for malignant reasons, such as username, secret passwords and so on. This kind of website is known as website phishing. To identify and predict the phishing site, we have proposed an effective system which depends on the use of machine learning techniques. We execute Classification algorithms and techniques to extricate the phishing informational collections and organize their legitimacy. The phishing site can be characterized dependent on certain significant traits, for example, URL and Domain Names, and rules for confirmation and encryption in the last phishing discovery period. With the guide of this program customers can buy products online effectively decisively.***

Keywords: Data mining Algorithm, Feature extraction, Machine Learning, Phishing, Phishing Attacks, Phishing Detection, Phishing website

1. Introduction

Phishing is a fraudulent effort to gain confidential user details for harmful reasons such as username, password, bank account numbers, credit card data. Construction of a new website which is visually and semantically similar to the original website is quite easy.

Phishers use these places to compile the customers' touchy info. In addition to the aggressors, not many security inquiries are asked to respond, acting as a significant level of security effort for the clients. As consumers react to these inquiries they get swept up in phishing assaults. Currently, phishing may be the most widely used cybercrime today. Because most consumers go online to get to the administrations that the government or banks have issued, there has been a dramatic rise in phishing for as far back as barely years. Phishers were taking in currency, so they worked out how to do as that as a profitable organization. Phishing forms include clone phishing, mobile phishing, DNS phishing and considerably more. In the off probability of creativity continuing to evolve, phishing techniques are quickly progressing, and the usage of phishing apparatuses aggressive to differentiate phishing would discourage this. As of late, AI is one of the most notable weapons available to counter phishing assaults.

Phishing attacks are generally broken down into four categories: Phishing, Spear Phishing, Replica Phishing, Whaling.

Phishing: A phishing technique where an intruder imitates a reliable individual to get mystery information, for instance, usernames, passwords, record number, etc. A continuous event is that couple of phishing sends have been professed to have been sent from the American Express, anyway they have not sent it.

Spear Phishing: A form of phishing involving parodying emails that are sent to a man or organization. In a typical phishing attack, the phished messages are submitted to an unusual emailId or archive when the messages derive in stick phishing from a conducted beneficiary.

Some of the cases where the lance phishing attack took place and centered on the RSA protection company, where the aggressors delivered phishing messages to four different RSA parent agency members.

Clone Phishing: A phishing technique that mimics a legitimate email account using an actual email and changes in the connection.

Whaling: A methodology of phishing that celebrates goals by people including government officials, big names and administrators. This is regarded as the most harmful kind of phishing in which the content of the email includes customers' complaints, official concerns and so on.

2. Literature Review/Survey

Phishing has become one of the major issue in recent times. Phishing generally occurs when the user clicks the URL of the phishing websites and enter his details. The phisher designs his website which is same as the original websites visually, one can rarely identify the phishing website by looking at it.

Types of Phishing Detection:
1) Phishing Detection based on Content: This approach tests the content on the Webpage to assess if the Webpage is phished. This uses the TF-IDF (Term Frequency-Inverse Record Frequency) based data recovery measure. When a website is provided, the formula ascertains the TD IDF scores on each and every web page. It utilizes the heuristics to decide if the site is phished.

2) Detection of Phishing based on Visual: process, the web pages are turned into low-target pictures using indicators such as hues, instructions for creating a picture signature. This method worked well, based on observations by the researchers. It can struggle though if the intruder generates a phishing website different from the starting one.

3) Phishing Detection based on Identity: This methodology recognizes a phishing page & its objective utilizing the Semantic Link Network (SLN). This method is approved utilizing research databases like thousand genuine phishing website pages and thousand phishing site pages.

4) Detection of Phishing based on Features: website highlights are extricated using the terms pack process. To extricate the highlights, the URL is partitioned into three regions, e.g. convention, room and path.

The core objective of our project is to identify the phishing websites based on the URL.

Python has a huge collection of built-in standard libraries, we have used a few of them like numpy, matplot, pandas, seaborn and a few others.

Our Dataset consists of 30 features (like URL length, page traffic, double slash for redirecting) identifying the phishing websites. The dataset contains different values 1,0 and -1 indicating the risk of opening the website. '1' indicates that the website is legitimate, '0' indicates that the website is suspicious(it may be safe website or a legitimate one) and '-1' indicates the website is safe and you can browse in the website safely.1 in the URL_Length column indicates that the website is identified as phishing website because of the length of the URL. Considering all the columns we compute the Result column. This column gives us the final result of a website whether it is a safe website or legitimate website. This column doesn't contain the value 0.

Firstly, we found the correlation between every two features. As we know, if the value of correlation with target variable is nearby zero then the effect of the variable on target is negligible. Therefore we removed a few columns of which the correlation with target variable is between -0.03 and +0.03 and the number of columns comes down to twenty six. Around Seventy percent of the preprocessed data is used to train the model. We used different models namely Random Forest Classifier, Logistic Regression and Support Vector Machine to train the model and compared the accuracy. Highest accuracy is achieved through Random Forest Classifier model. Then we test the model using the remaining thirty percent of data.

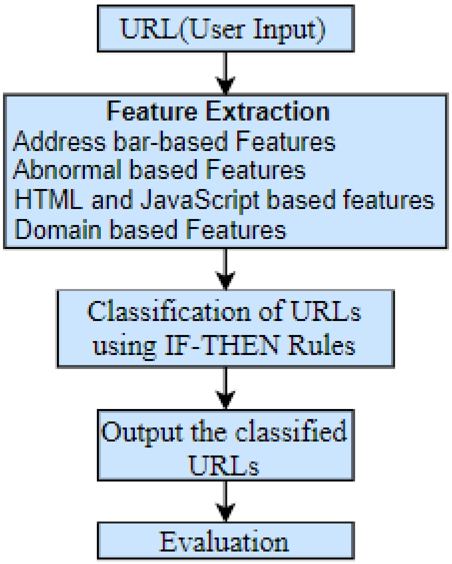
3. Existing System

In the previous models, phishing websites are detected using the bag-of-words method. This method initially converts the raw data to the vector of numbers and gives the vector as an input to the model. In general, the bag-of-words is a straight-forward and uncomplicated way to extract the features of data and use it in modelling. It involves several steps, firstly we collect the data and design the vocabulary i.e., we identify the unique words or patterns in the URL. Next step involves creating a vector. We identify the number of times the specified pattern occurred in the URL and create a vector of numbers based on the data. We can also identify N-grams to create the vector. N-grams is joining the two or more patterns and identifying their collective occurrences.

4. Proposed System

After pre-processing the data, the data set containing the information about legitimate websites and phishing websites is provided as an input for the program. The dataset includes 30 features of about 11000 websites that are used to differentiate phishing websites from legitimate ones. Each category has its own phishing attribute characteristics, and values are described. For each URL the specified characteristics are extracted, and valid input ranges are defined. Then these values are applied to the danger of each Phishing website. We give values for each input -1 for phishing websites,0 for suspect websites and 1 for legitimate websites, while the result consists of -1 and 1 representing the phishing website and a safe website respectively. After this information is prepared, we apply not many AI calculations to the dataset and figure out which calculation gives most noteworthy precision.

5. Architecture

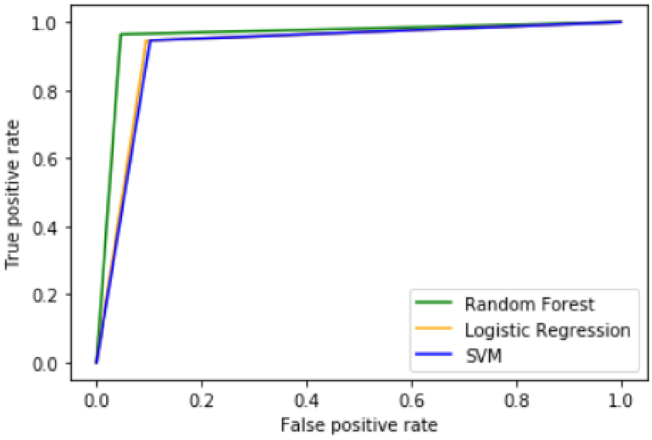


If From the above figure we can see the structure of client's search history arrangement. It describes how our framework is organized and in a theoretical view outlines the connection between different components in the framework. The measures utilized in the gadget design found in the above figure are as per the following

1) User Input: Firstly, the user supplies the URL or the document containing the bunch of URLs not knowing whether they are safe to browse or not. The system gathers information from the number of customers, who have used the website previously and it identifies the highlights of the behavior of the given URL . These highlights provide a short portrayal on the type of URL while extending the reaction time of the gadget.
2) Feature Extraction: From all the related highlights of the URL which are used to differentiate phishing URLs from the safe one's are eliminated at this stage. The URL function is partitioned into several types, such as Address bar-dependent highlights, intermittent highlights, HTML and JavaScript-dependent highlights, and domain dependent highlights.
3) URL Classification: The highlights gained from the preceding advance depend on various heuristics. Total number of heuristics that are used to determine the phishing, dubious or lawful existence of a URL is approximately sixteen. Given the heuristics given and the highlights removed, those highlights are added to the proposed rules to arrange a URL. 4 . Output the classified URLs: The outcomes are shown on the User Interface as diagrams or tables or graphs which gives us a pictorial portrayal of the generated outcomes. Likewise, for every URL on a size of 1-5 the extent of the phishing is learned to evaluate the impact of phishing on that URL.
4) Evaluating the Result: Generally, the impacts of

classification are calculated based on the values of Precision and Recall. A risk grid is plotted and it is usually used to determine the correctness of the classification technique by calculating the True Positives, Fake Positives, Fake Negatives and Actual Negatives.

6. Experimental Analysis



Our Approach considers Support-Vector Machine(SVM), Logistic regression and Random Forest Classification(RFC) for predicting the results. The below figure shows the Receiver Operating Characteristic Curve (ROC Curve) for all the three classification models. As we can see Random Forest Classification model gives the highest accuracy compared to the other two. This is because there are many decision tress involved in this model resulting to less over-fitting and least error.

7. Conclusion

Because phishing attacks are considered to be very dangerous and it is crucial that we have a system to detect them. When very sensitive and confidential user details can be leaked by phishing websites, this topic is more crucial to tackle. This problem can be solved simply by using the classifier for either of the machine learning algorithms. We already have classifiers that offer strong predictive phishing rates besides, but after our study it would be simpler and easier to use mixture of methodologies for forecasting and further increment the predictive accuracy rate of phishing websites. We observed that the existing framework gives less accuracy so we've proposed another phishing strategy that utilizes URL-based features, and we've also generated classifiers through several techniques of machine learning.

8. Future Scope

Later on we will demonstrate progress in the off chance we have a simplified phishing dataset from any other technique. Later on we will use at least two classifiers with a combination of the other to reach optimal accuracy. Additionally, we want to study different phishing techniques using Lexical highlights, organize focused highlights, content-related highlights, website-based highlights, and HTML and JavaScript website specific software that can help implement gadgets. Actually, we get highlights from URLs, then pass them through the different classifiers.

9. Acknowledgment

The authors wish to thank Department of CSE, SNIST for providing infrastructure for our project.

References

[1] Wong, R. K. K. (2019). “An Empirical Study on Performance Server Analysis and URL Phishing Prevention to Improve System Management Through Machine Learning”. In Economics of Grids, Clouds, Systems, and Services: 15th International Conference, GECON 2018, Pisa, Italy, September 18-20, 2018, Proceedings (Vol. 11113, p. 199). Springer.

[2] Rao, R. S., & Pais, A. R. (2019). Jail-Phish: “An improved search engine based phishing detection system”. Computers & Security, 83, 246-267.

[3] Ding, Y., Luktarhan, N., Li, K., & Slamu, W. (2019). “A keyword-based combination approach for detecting phishing webpages”. computers & security, 84, 256-275.

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

[5] Shekokar, N. M., Shah, C., Mahajan, M., & Rachh, S. (2015). “An ideal approach for detection and prevention of phishing attacks”. Procedia Computer Science, 49, 82-91.

[6] Rathod, J., & Nandy, D. “Anti-Phishing Technique to Detect URL Obfuscation”.

[7] Hodi, A., Kevri, J., & Karadag, A. (2016). “Comparison of machine learning techniques in phishing website classification”. In International Conference on Economic and Social Studies (ICESoS'16) (pp. 249-256).

[8] Pujara, P., & Chaudhari, M. B. (2018). “Phishing Website Detection using Machine Learning: A Review”.

[9] Desai, A., Jatakia, J., Naik, R., & Raul, N. (2017, May). “Malicious web content detection using machine learning”. In 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 1432-1436). IEEE.

[10] Lakshmi, V. S., & Vijaya, M. S. (2012). “Efficient prediction of phishing websites using supervised learning algorithms”. Procedia Engineering, 30, 798-805.

[11] Jain, A. K., & Gupta, B. B. (2018). “PHISH-SAFE: URL features-based phishing detection system using machine learning”. In Cyber Security (pp. 467-474). Springer, Singapore.

[12] Kazemian, H. B., & Ahmed, S. (2015). “Comparisons of machine learning techniques for detecting malicious webpages”. Expert Systems with Applications, 42(3), 1166-1177.

[13] Mao, J., Bian, J., Tian, W., Zhu, S., Wei, T., Li, A., & Liang, Z. (2019). “Phishing page detection via learning classifiers from page layout feature”. EURASIP Journal on Wireless Communications and Networking, 2019(1), 43.

[14] Mohammad, R. M., Thabtah, F., & McCluskey, L. (2012, December). “An assessment of features related to phishing websites using an automated technique”. In 2012 International Conference for Internet Technology and Secured Transactions (pp. 492-497), IEEE.
[15] Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). “Machine learning based phishing detection from URLs”. Expert Systems with Applications, 117, 345-357.
[16] Yuan, H., Chen, X., Li, Y., Yang, Z., & Liu, W. (2018, August). “Detecting Phishing Websites and Targets Based on URLs and Webpage Links”. In 2018 24th International Conference on Pattern Recognition (ICPR) (pp. 3669-3674). IEEE.