Web Phishing Detection

Abstract

This article surveys the literature on the detection of phishing attacks. Phishing attacks target vulnerabilities that exist in systems due to the human factor. Many cyber-attacks are spread via mechanisms that exploit end users' weaknesses, making users the weakest element in the security chain. The phishing problem is broad and no single silver-bullet solution exists to mitigate all the vulnerabilities effectively, thus multiple techniques are often implemented to mitigate specific attacks. This paper aims at surveying many of the recently proposed phishing mitigation techniques. A high-level overview of various categories of phishing mitigation techniques is also presented, such as detection, offensive defense, correction, and prevention, which we believe is critical to present where the phishing detection techniques fit in the overall mitigation process.

Introduction

Problem Statement

Phishing site detection is truly an unpredictable and element issue including numerous components and criteria that are not stable. On account of the last and in addition ambiguities in arranging sites because of the intelligent procedures programmers are utilizing, some keen proactive strategies can be helpful and powerful tools can be utilized. Several conventional techniques for detecting phishing websites have been suggested to cope with this problem. However, detecting phishing websites is a challenging task, as most of these techniques are not able to make an accurate decision

Definition

The definition of phishing attacks is not consistent in the literature, which is because the phishing problem is broad and incorporates varying scenarios. For example, according to PhishTank1:

"*Phishing is a fraudulent attempt, usually made through email, to steal your personal information"*

PhishTank's definition holds in several scenarios which, roughly, cover the majority of phishing attacks (although no accurate studies have been made to reliably quantify this). However, the definition limits phishing attacks to steal personal information, which is not always the case.

For example, a socially engineered message can lure the victim to install a Man in the Browser (MITB) malware (e.g., in the forms of web browser ActiveX components, plugins, or email attachments) which would in turn transfer money to the attacker's bank account, whenever the victim logs in to perform his/her banking tasks, without the need to steal the victim's personal information.

Thus we consider that PhishTank's definition is not broad enough to encompass the whole phishing problem.

Project Description

Recent years have witnessed the increasing threat of phishing attacks on mobile computing platforms. Mobile phishing is particularly dangerous due to the hardware limitations of mobile devices and mobile user habits.

We have developed a website with a search engine to detect phishing websites. If the user enters the URL, it checks whether the website is a phishing website or not by using several algorithms and data science concepts.

The Methodology used in this solution

To detect and predict e-banking phishing websites, we proposed an intelligent, flexible and effective system that is based on using classification algorithms. We implemented classification algorithms and techniques to extract the phishing datasets criteria to classify their legitimacy. The e-banking phishing website can be detected based on some important characteristics like URL and domain identity, and security and encryption criteria in the final phishing detection rate. Once a user makes a transaction online when he makes a payment through an e-banking website our system will use a data mining algorithm to detect whether the e-banking website is a phishing website or not.

System Design

The proposed methodology imports a dataset of phishing and legitimate URLs from the database and the imported data is pre-processed. Detecting phishing websites is performed based on four categories of URL features: domain-based, address-based, abnormal-based, and HTML, JavaScript features. These URL features are extracted with processed data and values for each URL attribute are generated.

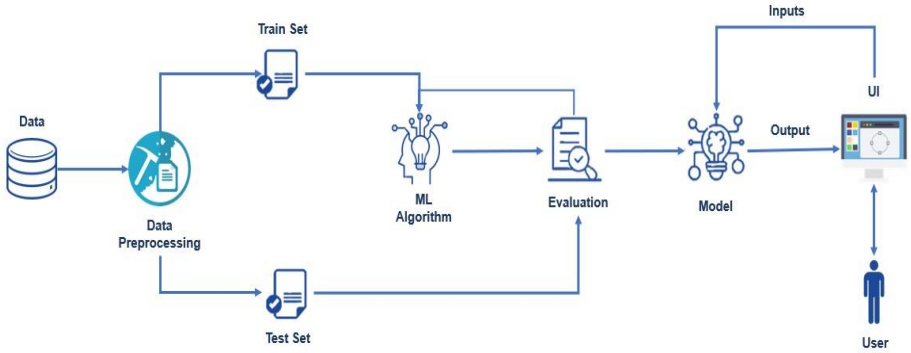
The analysis of URL is performed by machine learning technique which computes range value and the threshold value for URL attributes. Then it is classified into phishing and legitimate URL.

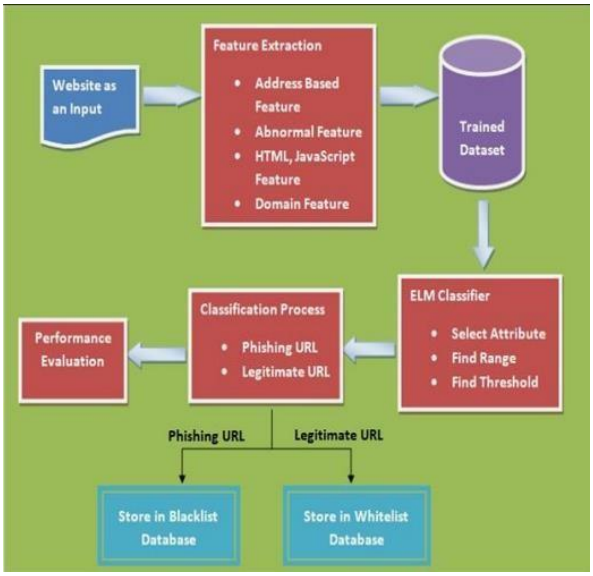
The attribute values are computed using feature extraction of phishing websites and it is used to identify the range value and threshold value.

The value for each phishing attribute is ranging from f-1, 0, and 1g these values are defined as low, medium, and high according to the phishing website feature. The classification of phishing and legitimate websites is based on the values of attributes extracted using four types of phishing categories and a machine learning approach.

System Architecture:

Overall Architecture:





MITIGATION OF PHISHING ATTACKS: AN OVERVIEW

Due to the broad nature of the phishing problem, we find it important to visualize the life-cycle of the phishing attacks, and based on that categorize anti-phishing solutions. Based on our review of the literature, we depict a flowchart describing the life-cycle of phishing campaigns from the perspective of anti-phishing techniques, which is intended to be the most comprehensive phishing solutions flowchart.

When a phishing campaign is started (e.g. by sending phishing emails to users), the first protection line is detecting the campaign. The detection techniques are broad and could incorporate techniques used by service providers to detect the attacks, end-user client software classification, and user awareness programs.

The ability to detect phishing campaigns can be enhanced whenever a phishing campaign is detected by learning from such an experience. For example, by learning from previous phishing campaigns, it is possible to enhance the detection of future phishing campaigns. Such learning can be performed by a human observer, or software (i.e. via a machine learning algorithm).

Once the phishing attack is detected, several actions could be applied against the campaign. According to our review of the literature, the following categories of approaches exist:

- Offensive defense — these approaches aim to attack phishing campaigns to render them less effective. This approach is particularly useful to protect users that have submitted their details to attackers.

- Correction — correction approaches mainly focus on taking down the phishing campaign. In the case of phishing websites, this is achieved by suspending the hosting account or removing phishing files.

- Prevention — phishing prevention methods are defined differently in the literature depending on the context. In this survey, the context is attempting to prevent attackers from starting phishing campaigns in the future.

However, if the phishing campaign is not detected (let it be detected by a human or a software classifier), then none of these actions can be applied. This emphasizes the importance of the detection phase.

Features Used for Phishing Domain Detection

There are a lot of algorithms and a wide variety of data types for phishing detection in the academic literature and commercial products. A phishing URL and the corresponding page have several features which can be differentiated from a malicious URL. For example; an attacker can register a long and confusing domain to hide the actual domain name **(**Cybersquatting, Typosquatting). In some cases, attackers can use direct IP addresses instead of using the domain name. This type of event is out of our scope, but it can be used for the same purpose. Attackers can also use short domain names that are irrelevant to legitimate brand names and don't have any FreeUrl addition. But these types of websites are also out of our scope because they are more relevant to fraudulent domains instead to phishing domains.

Besides URL-Based Features, different kinds of features which are used in machine learning algorithms in the detection process of academic studies are used. Features collected from academic studies for phishing domain detection with machine learning techniques are grouped as given below.

1. URL-Based Features

2. Domain-Based Features
3. Page-Based Features
4. Content-Based Features

URL-Based Features

URL is the first thing to analyze a website to decide whether it is phishing or not. As we mentioned before, URLs of phishing domains have some distinctive points. Features that are related to these points are obtained when the URL is processed. Some of the URL-Based Features are given below.

- Digit count in the URL
- Total length of URL
- Checking whether the URL is Typosquatted or not. (google.com → goggle.com)
- Checking whether it includes a legitimate brand name or not (apple iCloud-login.com)
- Number of subdomains in URL
- Is Top Level Domain (TLD) one of the most commonly used ones?

Domain-Based Features

The purpose of Phishing Domain Detection is to detect phishing domain names. Therefore, passive queries related to the domain name, which we want

to classify as phishing or not, provide useful information to us. Some useful Domain-Based Features are given below.

- Its domain name or its IP address in blacklists of well-known reputation services?
- How many days passed since the domain was registered?
- Is the registrant's name hidden?

Page-Based Features

Page-Based Features using information about pages which are calculated reputation ranking services. Some of these features give information about how reliable a website is. Some of the Page-Based Features are given below.

- Global Pagerank
- Country Pagerank
- Position at the Alexa Top 1 Million Site

Some Page-Based Features give us information about user activity on the target site. Some of these features are given below. Obtaining these types of features is not easy. There are some paid services for obtaining these types of features.

- Estimated Number of Visits for the domain on a daily, weekly, or monthly basis

- Average Pageviews per visit
- Average Visit Duration
- Web traffic share per country
- Count of references from Social Networks to the given domain
- Category of the domain
- Similar websites etc.

Content-Based Features

Obtaining these types of features requires an active scan to target the domain. Page contents are processed for us to detect whether the target domain is used for phishing or not. Some processed information about pages is given below.

- Page Titles
- Meta Tags
- Hidden Text
- Text in the Body
- Images etc.

By analyzing this information, we can gather information such as;

- Is it required to login into the website
- Website category
- Information about audience profile etc.

All of the features explained above are useful for phishing domain detection. In some cases, it may not be useful to use some of these, so there are some limitations to using these features. For example, it may not be logical to use some of the features such as Content-Based Features for the developing fast detection mechanism which can analyze the number of domains between 100.000 and 200.000. Another example would be if we want to analyze newly registered domains Page-Based Features is not very useful. Therefore, the features that will be used by the detection mechanism depend on the purpose of the detection mechanism. Which features to use in the detection mechanism should be selected carefully.

Detection Process

Detecting Phishing Domains is a classification problem, so it means we need labeled data that has samples as phish domains and legitimate domains in the training phase. The dataset which will be used in the training phase is a very important point to build a successful detection mechanism. We have to use samples whose classes are precisely known. So it means, that the samples which are labeled as phishing must be detected as phishing. Likewise, the samples which are labeled as legitimate must be detected as legitimate. Otherwise, the system will not work correctly if we use samples that we are not sure about.

There are so many machine learning algorithms and each algorithm has its working mechanism.

Decision Tree Algorithm

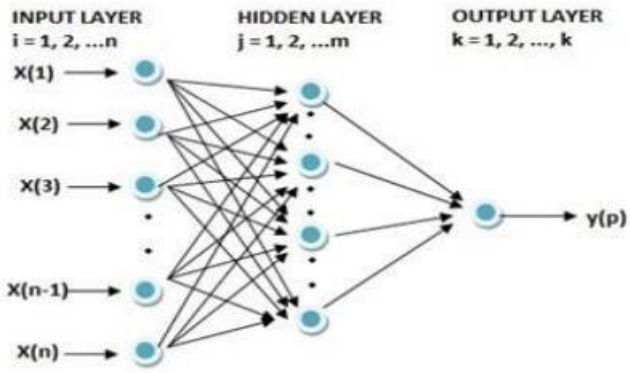
The Decision Tree Algorithm is a simple and powerful one. A Decision Tree can be considered as an improved nested-if-else structure. Each feature will be checked one by one.

Decision Tree uses an information gain measure that indicates how well a given feature separates the training examples according to their target classification. The name of the method is **Information Gain**. The mathematical equation of the information gain method is given below.

$$Gain(S, A) = \underbrace{Entropy(S)}_{\text{original entropy of S}} - \underbrace{\sum_{v \in values(A)} \frac{|S_v|}{|S|} \cdot Entropy(S_v)}_{\text{relative entropy of S}}$$

Extreme Learning Machine (ELM)

Extreme Learning Machine (ELM) is proposed as a single hidden layer feedforward artificial neural network (ANN) model which ensures high-performing learning and parameters such as threshold value, weight, and activation the function must have appropriate values for the data system which is to be modeled. In ELM learning, the parameters are gradient-based, where the input weights are randomly selected while the output weights are analytically calculated. For the sake of activating the cells in the hidden layer of ELM, a linear function as well as non-linear (sinus, sigmoid, Gaussian), and non-derivable or discrete activation functions can be used.



Conclusion

Systems varying from data entry to information processing applications can be made through websites. The entered information can be processed; the processed information can be obtained as output. Nowadays, websites are used in many fields such as scientific, technical, business, education, economy, etc.

Because of this intensive use, it can be also used as a tool by hackers for malicious purposes. One of the malicious purposes emerges as a phishing attack. A website or a web page can be imitated by phishing attacks and using various methods. Some information such as user's credit card information, and identity information can be obtained with these fake websites or web pages.

The purpose of the application is to make a classification for the determination of one of the types of attacks that cyber threats called phishing. Extreme Learning Machine is used for this purpose. In this study, we used a data set taken from the UCI website. In this dataset, input attributes are determined in 30, and the output attribute is determined in 1. Input attributes can take 3 different values which are 1, 0, and -1. Output attribute can take 2 different values which are 1, and -1. As a result of the study, the average classification accuracy was measured to be 95.34%.

REFERENCES

1. Zuochao Dou, Issa Khalil, Abdallah Khreishah, Ala AlFuqaha, and Mohsen Guizani, “Systematization of Knowledge (SoK): A Systematic Review of Software Based Web Phishing Detection”, Communications Surveys & Tutorials,2017.

2. Peng Yang, Guangzhen Zhao, Peng Zeng, “Phishing Website Detection based on Multidimensional Features driven by Deep Learning”, IEEE Access, 2018.

3. Christopher N. Gutierrezy, Taegyu Kimy, Raffaele Della Cortez, Jeffrey Averyyx, Dan Goldwassery, Marcello Cinquez, Saurabh Bagchiy, “Learning from the Ones that Got Away: Detecting New Forms of Phishing Attacks”, Transactions on Dependable and Secure Computing, 2018

4. W. D. Yu, S. Nargundkar, and N. Tiruthani, “A phishing vulnerability analysis of web-based systems,” in Proceedings of the 13th IEEE Symposium on Computers and Communications (ISCC 2008). Marrakech, Morocco: IEEE, July 2008, pp. 326–331

5. A. Y. Fu, L. Wenyin, and X. Deng, “Detecting phishing web pages with visual similarity assessment based on earth mover’s distance (emd),” IEEE Trans. Dependable Secure. Comput., vol. 3, no. 4, pp. 301–311, Oct. 2006.