Chapter 2

Review of Literature

In order to protect the users and organizations from phishing attacks, many anti-phishing solutions have been proposed which can be categorized as user education or software-based anti-phishing techniques. Moreover, Software-based techniques can be later divided as List based, Visual Similarity based, Content based and Heuristic based techniques.

List based anti-phishing techniques keep a black-list, white-list, or mix of both. Blacklist is a list of URLs or domains of known phishing websites published by a trusted party. Whenever a user enters URL in the address bar, the browser refers to the blacklist to check if the entered URL or domain exists within the blacklist. If the website exists within the blacklist, then the browser sends an alert message to the user. Similarly, the whitelist is the list of trusted URLs or domains. In this case, the user is warned if the entered URL doesnt exist in the whitelist. McAfees anti-phishing filter uses blacklist and PhishZoo depends on a whitelist for phishing website detection.

In Visual Similarity approach, the layout of the spoofed website and original website are compared by measuring the three metrics - layout similarity, block-level similarity and overall style similarity. The negative aspect of the list and visual similarity based approaches is that it usually doesnt cover newly launched (zero-hour attack) phishing websites.

Content based approach examines the content of a web page to decide if it is genuine or fake. This is in contrast to other approaches that only take a shallow view of the surface characteristics of a web page. This method has sufficient accuracy and low false alarms in determining the fake web page. Heuristic based approach uses common features of phishing and legitimate sites based on URL, Lookup, Search Engine, HTML DOM, Certificate and Website traffic. The websites are declared as phishing sites if the heuristic design of the websites matches the predefined rules.

There is one more technique that is Machine Learning approach, which not only verifies exhaustive features of the URLs as per heuristic design but also refers to large datasets in order to eliminate such cases where in the heuristic design solely is incapable to pass the correct verdict. Some of the works in the field of phishing detection and protection proposed by other authors in previous years are explained below.

Ankit Kumar Jain et al [2] presented a comprehensive analysis of all the Phishing attacks known along with their resulting consequences. Moreover, it also provides a very useful insight over the various machine learning based approaches for phishing detection with the help of a comparative study. This opens up various viewpoints in terms of finding more efficient solutions with help of machine learning in near future.

A detection technique for phishing websites was proposed by Abdulghani Ali Ahmed et al [1] which examined Uniform Resources Locators (URLs) of suspected web pages as per five extracted features. Phishtank and Yahoo directory datasets are used to assess the accuracy of the results given by proposed solution. The final report thus establishes that the detection mechanism can detect various types of phishing attacks without fail. However, there are still chances of receiving false alarms.

An algorithm which performs Googles updated blacklist check utilizes Google search engine results, Alexa Ranking and no of URL- based features, for detecting phishing URLs was suggested by Varsharani Ramdas Hawanna et al [3]. It displays an alert message if the URL classified as phishing, otherwise it displays a safe message. This algorithm enhances the performance when dealing with known/old phishing URLs.

Priyanka Singh et al [4] have executed two algorithms named Backpropagation network with Support Vector Machine (SVM) and Adaline network to improve the detection rate and classification using datasets of Phishtank and Alexa. Training time, testing time, mean square error and prediction accuracy are the parameters used to evaluate the performance of both algorithms. Adaline network gave 99.14% accuracy. Training time used by the Adaline network is very less when compared with the Backpropagation network with SVM.

A Hybrid Model based approach which used 30 features to solve the phishing websites problem was presented by Sohail Asghar et al [5]. A single model cannot efficiently detect the phishing websites, therefore to enhance the accuracy, efficiency and performance rate, two or more models are combined together to form a more robust classifier. Firstly, the individual performance of a classifier is checked and then the best classifier in terms of high accuracy and less error rate is evaluated. The best classifier model is combined with other classifiers one by one and finally, a better hybrid classification model is achieved.