Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 2, February 2014, pg.696 – 699

REVIEW ARTICLE

PHISHING WEBSITE DETECTION: A REVIEW

Feon Jaison[1], Seenia Francis[2]

[1]Department of Computer Science & Engineering, Jyothi Engineering College, Cheruthuruthy
[2]Assistant Professor, Department of Computer Science & Engineering, Jyothi Engineering College, Cheruthuruthy

Abstract—Phishing is an attempt to steal users' personal and financial information such as passwords, credit card numbers, through electronic communication such as e-mail and other messaging services. Attackers pretend to be from a organization which direct the users to a fake website that resembles a phishing website, which is then used to collect users personal information. Attackers can also trick users into downloading malicious codes or malware after they click on a link embedded in the email.

Various researches have been done for protecting the users from phishing attacks. They include firewalls, blacklisting certain domains and internet protocol (IP) addresses, spam filtering techniques, fake website detection, client side tool-bars and user education. Each of these existing techniques has some advantages and some disadvantages. The need to automatically discover a phishing target is an important problem for anti-phishing efforts. If we know the webpage which is considered as the target webpage, we can confirm which all are the phishing pages. It could help the owners to identify phishing attacks so that they can immediately take necessary counter measures.

Keywords-Phishing Website, Division Clustering Algorithm, Classifiers

I. INTRODUCTION

The paper discuss with the importance of Phishing Detection Websites[5] and a review of various models in detection. Phishing attacks are significant threat to users of the Internet causing tremendous loss year by. The goal here is to combine the best aspects of human verified blacklists and heuristic-based methods which are have the low false positive rate of the owner. The key insight behind our detection algorithm is to define the existing human-verified blacklists and apply various techniques. The features introduced in Carnegie Mellon Anti-Phishing and Network Analysis Tool (CANTINA)[3], in similarity feature to a machine learning based phishing detection. The heuristic detection model mainly makes the use of various characteristics of the URL that includes URL similarity calculation, domain name probability evaluation, IP address, the port number, etc. It will get the information of website ranking, registration information, category in which phishing websites and other information by querying the third party libraries such as Google Page rank. This will increase the phishing detection efficiency compared to older signature based models. The method can detect the various websites containing phishing attacks and abnormal behaviors.

Web content, which is the main display channels for phishing fraud, well expresses the various intention of the website. The different machine learning module, Decision Tree, Support Vector Machine, Naive Bayes, Neural Network and other machine learning algorithm have been applied into the model training and predicting whether the given website is phishing website or not. Clustering methods and classification are both fundamental tasks in Data Mining. Classification is used mostly as a supervised learning method, clustering for unsupervised learning. The goal of clustering is descriptive, that of classification is predictive. Clustering groups data instances into subsets in such a manner that similar instances are grouped together, while different instances belong to different groups. The instances are thereby categorized into an efficient representation. Formally, the clustering structure is represented as a set of sub-sets. Distance of clustering is the grouping of similar instances/objects. There are two main type of measures used to estimate this relation: distance measures and similarity measures. Many clustering methods use distance measures to determine the similarity or dissimilarity between any pair of objects. It is useful to denote the distance between two instances xi and xj as d(xi,xj).

Similarity Functions is considered as an alternative concept to that of the distance is the similarity function s(xi, xj) that compares the two vectors xi and xj. This function should be symmetrical i.e s(xi,xj)=s(xj,xi) and have a large value when xi and xj are somehow similar and constitute the largest value for identical vectors. Cosine Measure is the angle between the two vectors is a meaningful measure of their similarity, the normalized inner product may be an appropriate similarity measure: The existing phishing models describes as Blacklist/White list based method in which a user visits a Web site so that anti-phishing tool searches the address of that site in a blacklist stored in the database.

$$s(x_i, x_j) = \frac{x_i^T \cdot x_j}{\|x_i\| \cdot \|x_j\|}$$

If the visited site is present on the list, the anti-phishing tool will instruct the users. Tools in this category include Scam Blocker from the EarthLink Company, Phish Guard, and Net craft, etc. In the Rule-based method the various tools uses certain rules in their software, and checks the security of a Web site according to these rules. Examples of this type of tools include Spoof Guard developed by Stanford, Trust Watch of the Geo Trust, etc. Spoof Guard checks the domain name, URL of Web site, it also checks whether the browser is directed to the current URL through the links in the contents of e-mails. If it identifies that the domain name of the visited Web site is similar to a well-known domain name, or if they are not using the standard port, Spoof Guard will instruct the users about the phishing website. Both Spoof Guard and Trust Watch provide a toolbar in the browsers to notify their users whether the Web site is verified and trusted from phishing.

The Intelligent architecture for the phishing website detection is that Feature Extraction module Heterogeneous Classifier that are built from the features extracted, Ensemble Classification Process, Hierarchical Clustering Algorithm for categorizing the various phishing websites.

1) Feature selection [1] is an important problem for pattern classification systems. The process includes how to select good features according to the maximal statistical dependency criterion based on mutual information. Because of the difficulty in directly implementing the maximal dependency condition, the procedure derives an equivalent form, called minimal redundancy-maximal-relevance criterion, for first-order incremental feature selection. The process basically include selecting features from N samples .i.e. the feature selection problem is to find from M-dimensional observation space R(m),a subspace of m features. Selecting the features with highest relevance to target class C is de-fined as Max-relevance.

$$I(x;y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy.$$

2) A naive Bayes classifier[4] considers all of these properties to independently contribute to the probability that this fruit is an apple. It only requires a small amount of training data to estimate the parameters necessary for classification, i.e. mean and variance The naive Bayes probabilistic model. The probability model for a classifier is a conditional model over a dependent class variable C with a small number of classes, conditional on several feature variables F1 through Fn is given by.

$$p(C|F_1,\dots,F_n)$$

3) The best way to define statistical learning is called as supervised Learning. In this each data points consist of a vector of features denoted as x and a class label y, and it is assumed that there is some underlying function f such that y=f(x) for each training data point (x,y).The goal of learning algorithm is to find a good approximation h to f that can be applied to assign labels to new x values. The function h is called a classifier, because it assigns class labels y to input data points x. supervised learning can be applied to many problems including hand-writing recognition, medical diagnosis. Ordinary machine learning algorithms work by searching through a space of possible functions that is called as the hypothesis. To find one function h which is the best approximation to a unknown function f we uses the ensemble classifier process. To determine which hypothesis h is the best, a learning algorithm can measure how well h matches f on the training data points, and it can also assess how consistent h is with any variable prior to knowledge about the existing problem.

4) A clustering [2] is the data mining technique used to place data elements into related groups. The data mining analyze data in different perspective and classifies the data and summarize it into useful information. Analyzing is done using cluster analysis, Induction, Decision tree etc. This most popular methods to separate data into disjoint groups. Hierarchical clustering will measure the distance between 2 tuples and which specifies the dissimilarity in the sets as a function of the pair-wise distance.

II. APPLICATIONS

The 4 different ways in which the antiphishing software can be deployed is that.

1) Web browsers have integrated an antiphishing filter into the browser itself.
2) At least one brand of security software has integrated an anti-phishing filter into its antivirus program and its Internet security software.
3) There is anti-phishing software available from at least one company specifically for routers.
4) Email software may include an antiphishing filter, or email blocking may be offered by a web host. McAfee Internet Security include antivirus and antiphishing services

III. CONCLUSION

Phishing has becoming a serious network security problem, causing financial loss of billions of dollars to both consumers and e-commerce companies. Phishing attacks can be detected through a combination of customer reportage, bounce monitoring, image use monitoring, honey pots and other techniques. Email authentication technologies such as Sender-ID and cryptographic signing, when widely deployed, have the potential to prevent phishing emails from reaching users. Personally identifiable information should be included in all email communications. Systems allowing the user to enter or select customized text and imagery are particularly promising. Anti-phishing toolbars are promising tools for identifying phishing sites and heightening security when a potential phishing site is detected. By IPDCM it includes the detection of phishing websites through ensemble classifiers and categorizing the phishing websites according to the various streams as online payments, Banking etc.

REFERENCES

1) Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max- Relevance, and Min-Redundancy Hanchuan Peng, Member, IEEE, Fuhui Long, and Chris Ding.
2) Hierarchical Clustering Algorithm - A Comparative Study,Dr.N.Rajalingam K.Ranjini Dept. of Management Studies.
3) CANTINA: A Content-Based Approach to Detecting PhishingWeb Sites.Yue Zhang Dept of Computer Science University of Pittsburgh 210 South Bouquet Street.
4) Naive Bayes Classifier.
5) Web Phishing Detection In Machine Learning Using Heuristic Image Based Method, Vinnarasi Tharania. I, R. Sangareswari , M. Saleembabu International Journal of Engineering Research and Applications ISSN: 2248-9622