Web Phishing detection

Introduction:

The purpose or goal behind phishing is data, money or personal information stealing through the fake website. The best strategy for avoiding the contact with the phishing web site is to detect real time malicious URL. Phishing websites can be determined on the basis of their domains. They usually are related to URL which needs to be registered (low-level domain and upper-level domain, path, query). Recently acquired status of intra-URL relationship is used to evaluate it using distinctive properties extracted from words that compose a URL based on query data from various search engines such as Google and Yahoo. These properties are further led to the machine-learning based classification for the identification of phishing URLs from a real dataset. This paper focus on real time URL phishing against phishing content by using phish-STORM. For this a few relationships between the register domain rest of the URL are consider also intra URL relentless is consider which help to dusting wish between phishing or non-phishing URL. For detecting a phishing website certain typical blacklisted urls are used, but this technique is unproductive as the duration of phishing websites is very short

Literature review:

A new hybrid ensemble feature selection framework for machine learning-based phishing detection system:

In this paper, the authors proposed a system with a collection or set of Hybrid features to classify websites based on machine learning algorithms. The main feature set is extracted using the cumulative distribution gradient technique, while the data perturbation ensemble technique is used to extract the secondary feature set. The algorithm used for training the classifier is Random Forest in association with ensemble learner identifies the phishing websites with a precision of 94.6 percent.

Comparative Study of the Detection of Malicious URLs Using Shallow and Deep Networks:

The authors made a relative study to detect phishing website URLs with machine learning and deep learning algorithms. Convolution Neural Network (CNN) and CNN Long Short-Term Memory (CNN-LSTM) with Logistic Regression formed the architecture of the classification model. The system was designed using tools like TensorFlow along with Keras for machine learning and deep learning model. The dataset was imported from multiple sources to provide better scalability. The phishing website URL dataset was obtained from OpenPhish and Phishtank, while the malicious or spam website URLs were imported from MalwareDomains.

Anomaly based web phishing page detection:

The proposed system detected phishing websites using a machine learning algorithm. The feature set included six features based on the website structure and was chosen after a comparative study by the authors. The classifier was trained using Support Vector Machine which worked effectively to classify websites whether legitimate or phishing. The model presented obtained an accuracy of 84 percent for the classification of websites.

Malicious web content detection using machine learning:

In this paper, the authors designed a browser extension to detect phishing websites. The system used multiple machine learning algorithms which included Random Forest, Support Vector Machine (SVM), and k-Nearest Neighbor (kNN) to train the classifier to achieve higher precision by doing a comparative study. The feature set included a content-based approach for extracting the JavaScript and HTML features of the websites. The dataset was imported from UCI-Machine Learning Repository and boasted a 22 feature classification technique to detect phishing websites.

An examination of machine learning systems for phishing recognition:

Authors made a comparative study of various machine learning algorithms such as Random Forests (RF), Support Vector Machines (SVM), Logistic Regression (LR), Bayesian Additive Regression Trees (BART), and Neural Networks to implement an efficient phishing website detection system. The dataset imported included a list of 2889 websites which were termed as phishing and a set of true blue messages. In total 43 features were extracted from the acquired dataset and were used extensively to train the classifier using the machine learning algorithms to obtain higher precision and accuracy.

On Feature Selection for the Prediction of Phishing Websites:

This paper proposes a phishing website detection method using reduces feature classification. The extracted features were analyzed using Support Vector Machine (SVM) and Logistic Regression algorithms. Out of the total 30 features identified, 19 features were selected and used for classification. The model was implemented using Big Data and the Dataset was obtained from the UCI Irvine machine learning repository. Between the two algorithms used, Support Vector Machine (SVM) showed better performance and accuracy of 95.62%.

Protecting user against phishing using Antiphishing:

AntiPhish is used to avoid users from using fraudulent web sites which in turn may lead to phishing attack. Here, AntiPhish traces the sensitive information to be filled by the user and alerts the user whenever he/she is attempting to share his/her information to a untrusted web site. The much effective elucidation for this is cultivating the users to approach only for trusted websites. However, this approach is unrealistic. Anyhow, the user may get tricked. Hence, it becomes mandatory for the associates to present such explanations to overcome the problem of phishing. Widely accepted alternatives are based on the creepy websites for the identification of “clones” and maintenance of records of phishing websites which are in hit list.

Learning to Detect Phishing Emails:

An alternative for detecting these attacks is a relevant process of reliability of machine on a trait intended for the reflection of the besieged deception of user by means of electronic communication. This approach can be used in the detection of phishing websites, or the text messages sent through emails that are used for trapping the victims. Approximately, 800 phishing mails and 7,000 nonphishing mails are traced till date and are detected accurately over 95% of them along with the categorization on the basis of 0.09% of the genuine emails. We can just wrap up with the methods for identifying the deception, along with the progressing nature of attacks [3].

Phishing detection system for e-banking using fuzzy data mining:

Phishing websites, mainly used for e-banking services, are very complex and dynamic to be identified and classified. Due to the involvement of various ambiguities in the detection, certain crucial data mining techniques may prove an effective means in keeping the e-commerce websites safe since it deals with considering various quality factors rather than exact values. In this paper, an effective approach to overcome the “fuzziness” in the e-banking phishing website assessment is used an intelligent resilient and effective model for detecting e-banking phishing websites is put forth. The applied model is based on fuzzy logics along with data mining algorithms to consider various effective factors of the e-banking phishing website.

CONCLUSION:

Phishing URL detection plays a pivotal role for many cybersecurity software and applications. In this paper, we researched and reviewed works based on the advanced machine learning techniques and approaches that promise a fresh approach in this domain. This article includes summary of the reviewed works after a systematic and comprehensive study on Phishing Website Detection systems. We believe that the presented survey would help researchers and developers with the insight of the progress achieved in the past years. Despite the tremendous progress in the field of cybersecurity, phishing website detection still pose a challenging problem with the ever evolving technology and techniques.

Reference:

[1]. “Protecting Users Against Phishing Attacks with AntiPhish” Engin Kirda and Christopher Kruegel Technical University of Vienna

[2]. “Learning to Detect Phishing Emails” Ian Fette School of Computer Science Carnegie Mellon University Pittsburgh, PA, 15213, USA icf@cs.cmu.edu Norman Sadeh School of Computer Science Carnegie Mellon University Pittsburgh, PA, 15213, USA Anthony Tomasic School of Computer Science Carnegie Mellon University Pittsburgh, PA, 15213, USA

[3]. Modeling and Preventing Phishing Attacks by Markus Jakobsson, Phishing detection system for e -banking using fuzzy data mining by Aburrous, M. ; Dept. of Comput., Univ. of Bradford, Bradford, UK ; Hossain, M.A. ; Dahal, K. ; Thabatah, F.

[4] M. Chandrasekaran, et al., ”Phishing email detection based on structural properties”, in New York State Cyber Security Conference (NYS) , Albany, NY ,” 2006

[5] P. R. a. D. L. Ganger, ”Gone phishing: Evaluating anti-phishing tools for windows. Technical report, ,” September 2006

[6] M. Bazarganigilani, ”Phishing E-Mail Detection Using Ontology Concept and Nave Bayes Algorithm,” International Journal of Research and Reviews in Computer Science, vol. 2,no.2, 2011.

[7] M. Chandrasekaran, et al., ”Phoney: Mimicking user response to detect phishing attacks,” in In: Symposium on World of Wireless, Mobile and Multimedia Networks, IEEE Computer Society, 2006, pp. 668-672

[8] I. Fette, et al., ”Learning to detect phishing emails,” in Proc. 16th International World Wide Web Conference (WWW 2007), ACM Press, New York, NY, USA, May 2007, pp. 649-656

[9] A. Bergholz, et al., ”Improved phishing detection using model-based features,” in Proc. Conference on Email and Anti-Spam (CEAS). Mountain View Conf, CA, aug 2008

[10] L. Ma, et al.,”Detecting phishing emails using hybrid features,”IEEE Conf, 2009, pp. 493-497