International Journal of Engineering& Scientific Research
Vol.9 Issue 9, September 2021,
ISSN: 2347-6532 Impact Factor: 6.660
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

A Survey on the Usage of Machine Learning Techniques for the Detection of Phishing Websites

Nagaraju T A
Assistant professor
Department of Electronics and Communication
Government Engineering College, Ramanagara, India.

***Abstract*—**Phishing theft is an attempt to steal personal information, such as credit card details, account number, social security number, etc. It is a deceptive method that uses social engineering and technology to convince the victim to provide personal information, usually for financial gain. There has been a spate of phishing scandals, and this has led to the risk of identity theft and financial loss. The discovery of a website for phishing scams is very important for online banking users and e-commerce users. Detection of phishing website is a difficult task. In this work, real time detection of phishing using machine learning technique.

Keywords—Phishing, Machine Learning Techniques

Introduction

The year 1991 was marked by dramatic changes in the use of the Internet. In 1991, when the Internet was made available for use, many businesses were turned into websites. Ecommerce has grown with this dramatic change in business trends. This change has provided an opportunity for electronic data exchange and electronic wallet transfers. But with the advent of ecommerce, a new trend came to fruition known as Cybercrimes. One of the Cybercrimes is Phishing.

Phishing attacks may be directed at individuals or organizations. In the case of personal attacks, the intent is to exploit their bank accounts or financial accounts and the goal is usually to gain money. In the case of corporate attacks, it may be to win over their competitors, to retaliate, or to inflict dishonesty on such companies by leading to the loss of their reputation, and so on.

In a typical Phishing attack, the target receives an email containing website links from a legitimate looking source. Clicking on these web links may take the target to a webpage or download malicious software. A website that opens by clicking a link may be a fake page, with its official appearance and sound. This webpage may prompt users to provide personal information and by clicking the submit button, this information will be forwarded to the attacker's email ID. The attacker may abuse the information for personal gain.

Using secure connections, installing anti-theft tools, and ignoring phishing emails are some of the existing methods against cybercrime. Among the new emerging strategies, data mining has become one of the most effective ways of categorizing websites into fraudulent identity theft websites or official websites. Website attributes can be used as features and these features can be used as an effective way to differentiate a website.

Machine learning is a form of artificial intelligence that gives computers the ability to read without precise programming. Machine learning focuses on the development of computer programs that can teach growth and change as they are presented with new data. The process of machine learning is similar to that of data mining. Machine learning activities are usually divided into three broad categories, depending on the type of signal or response found in the learning program. Supervised learning: The computer is provided with model inputs and desired outputs, provided by the teacher, and the goal is to read a standard rule that displays outputs. Unsupervised learning: No labels are provided with the learning algorithm, which leaves it alone to determine the structure of its input. Unsupervised learning can be the goal itself or the path to the end. Reinforcement learning: A computer program works with a flexible environment in which it has to make a goal, unless the teacher clearly tells it that it is close to its goal. Between supervised and non-supervised learning is less supervised learning, in which the teacher gives an incomplete training signal: a set of training with some outcomes that are not available.

Machine learning and data mining often use similar methods and are very complex. They can be categorized as follows: Machine learning focuses on speculation, based on known facts learned from training data while data mining focuses on the acquisition of unknown data. This is a step-by-step analysis of Access to Information on Database.

PHISHING ATTACKS TYPES

Deceptive Phishing: The most popular type of fraudulent identity theft strategy, it refers to any attack where fraudsters impersonate a real organization and try to take private information or login credentials. Those messages use risks and a sense of loyalty to motivate clients to do what the attackers are doing.

Spear Phishing: The criminal tactics for stealing sensitive information with a spear, fraudsters recreate their attack messages by name, position, organization, job phone number and other data trying to trick the victim into believing he or she has a relationship with the sender. The purpose is the same as the crime of stealing sensitive information: draw the deceased by tapping the malicious URL or email link, with the goal that they will provide their information.

CEO Fraud: Phishing scammers use email sent by a professional person to request installments or information from others within the organization. The basic premise is: For the victim to exchange money directly from cybercriminals.

Pharming: Fraudsters steal website domain name and use it to redirect visitors to a fraudulent site. The main goal is to prevent and steal online payments

Dropbox Phishing: Real-time emails from the drop-down box ask the user to click to protect account or download a shared account. the main purpose is to install a malware program on the victim's computer.

Google Docs Phishing: The message invites the victim to view the documents in Google Docs. The landing page is actually in Google Drive so it looks authentic, but entering your information will send it directly to fraudsters. The goal is to access your Google, which includes Gmail, Google Play and the android app.

LITERATURE SURVEY

Phishing attack is growing in such a way that it has become a global problem. Phishing scams use sensitive user information for malicious actions. Exploring these malicious activities has proposed a number of ways to combat the crime of identity theft.

Aljofey A et al. [1] CNN-based detection system for identifying a crime page for stealing sensitive information. Sequential pattern is used to find URLs. Existing research shows that CNN performance is better at retrieving images than text.

AlEroud A et. al, [2] A productive opposition network is used in the study to bypass the detection system. The Neural Network based discovery system can detect a faulty network view by reading the location.

Hong J. et al., [3] Lexical features used to identify criminal websites to steal sensitive information. Performance tests were based on a search-based database. Thus, there is no guarantee of the success of the URL detector with real-time URLs.

Abdulhamit Subasi et. al, [4], introduced a clever system for detecting cyber-attacks. Use different data mining techniques to determine the categories of websites: legal or criminal identity theft. Different dividers were used to create a clever system for the detection of a criminal website to steal sensitive information. The accuracy of the layout, the area under the curves of the receiver performance feature and the F-measure are used to evaluate the effectiveness of data mining techniques. The results showed that Random Forest performed very well among the classification methods by obtaining a high accuracy of 97.36%.

Hossein Shirazi et.al, [5], a framework, called Fresh-Phish, was created, to create current machine learning data for criminal web theft websites. Using the 30 different website features they inquired about using python, they built a large database with a label and analyzed a few machine learning sections against the database to determine which one was more accurate. They not only analyze the accuracy of the process, but also how long it takes to train the model.

Rao RS et. al, [6] Authors who use page attributes include logo, favicon, text and styles. The method used the server to update page features that reduced the performance of the identification system.

Jain A.K., et. al, [7] Both NB and SVM algorithms have been used to identify malicious websites. Both SVM and NB are slow learners and do not retain previous results in memory. Thus, the efficiency of the URL detector can be reduced.

Hung Le et al., [8] A deep learning URL detector is suggested. The authors argue that the method may generate data from the URL. A depth learning methods require more time to produce a result. In addition, it processes the URL and acts as a library to generate output.

Samuel Marchal et. al, [9] has proposed the limits of the crime of stealing sensitive information from the issues they face while creating a web page. As a result, the Off-the-Hook method, which uses a few notable features including high accuracy, product independence and good language autonomy, speed of decision, resilience to identity theft and strong evolution in criminal crime theft techniques, is used. Off-the-Hook is used as a full-featured client-side browser, which maintains user privacy. In addition, Off-the-Hook identifies the targeted web page for sensitive information theft attempts to impersonate and incorporates this target into its warning. We tested Off-the-Hook in two different user studies. Results show that users prefer Off-the-Hook alerts over Firefox alerts.

Applying Neural Networks techniques produce an internal structure in which the system can learn through examples, differentiating inputs that are different from excavation techniques. Gayathri. S [10] has suggested a category for Phishing websites for stealing sensitive information using advanced polynomial neural networks using genetic algorithm.

ZouFutai, Gang [11] focus on the behavior pattern of criminal websites for stealing sensitive information. They analyzed the actual IP flow from the ISP and proposed a discovery method based on Graph Mining with Belief Propagation. Tests have suggested that their algorithm has good accuracy and operational efficiency.

Weiwei Zhuang [12] proposed a model in the design and use of a clever model for finding criminal websites to steal sensitive information. They have released 10 different types of features such as title, keyword and link text information to represent the website. Various classifiers were then developed based on these different characteristics. The basic integrated classification algorithm is used to compile predicted results, derived from the detectors for the detection of identity theft. Hierarchical clustering technology has been used to differentiate the automated phishing classification.

Phishing is a duplicitous attempt to snip user's personal information. The rise in the number of cybercrime websites, users need to stay safe online. Therefore, there is a need to develop strategies to fight crime of identity theft using machine learning algorithms

CONCLUSION

The survey reviewed a number of software strategies against anti-phishing There are some of the key elements in measuring solutions for sensitive identity theft are the accuracy of detection in relation to the attack on sensitive identity theft. This is because criminal websites for stealing sensitive information are usually short-lived and zero-hour detection is important. A system with high false positives can cause more harm than good. In addition, end users will get into the habit of ignoring security alerts if the separator is usually faulty. The use of machine learning techniques is promising as it has led to very effective phishing classifiers

References

[1] Aljofey A, Jiang Q, Qu Q, Huang M, Niyigena JP. An effective phishing detection model based on character level convolutional neural network from URL. Electronics. 2020 Sep; 9(9):1514.

[2] AlEroud A, Karabatis G. Bypassing detection of URL-based phishing attacks using generative adversarial deep neural networks. In: Proceedings of the Sixth International Workshop on Security and Privacy Analytics 2020 Mar 16 (pp. 53–60).

[3] Hong J., Kim T., Liu J., Park N., Kim SW, "Phishing URL Detection with Lexical Features and Blacklisted Domains", Autonomous Secure Cyber Systems. Springer,2019.

[4] Abdulhamit Subasi, Esraa Molah, Fatin Almkallawi, Touseef J. Chaudhery, ―Intelligent Phishing Website Detection using Random Forest Classifier‖ in 2017 International conference on Electrical and Computing Technologies and Applications (ICECTA).

[5] Hossein Shirazi, Kyle Haefner, Indrakshi Ray, ―Fresh-Phish: A Framework for Auto-Detection of Phishing Websites‖ in 2017 IEEE International Conference on Information Reuse and Integration.

[6] Rao RS, Pais AR. Jail-Phish: An improved search engine based phishing detection system. Computers & Security. 2019 Jun 1; 83:246–67.

[7] Jain A.K., Gupta B.B. "PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning", Cyber Security. Advances in Intelligent Systems and Computing, vol. 729, 2018.

[8] Hung Le, Quang Pham, Doyen Sahoo, and Steven C.H. Hoi, "URLNet: Learning a URL Representation with Deep Learning for Malicious URL Detection", Conference'17, Washington, DC, USA, arXiv:1802.03162, July 2017.

[9] Samuel Marchal, Giovanni Armano, TommiGrondahl, KalleSaari, Nidhi Singh, and N. Asokan, ―Off-the-Hook: An Efficient and Usable Client-Side Phishing Prevention Application‖ in IEEE Transactions on Computer, VOL. 66, NO. 10, OCTOBER 2017.

[10] Gayathri. S, ―Phishing Websites Classifier using Polynomial Neural Networks in Genetic Algorithm‖ in 2017 4th International Conference on Signal Processing, Communications and Networking (ICSCN -2017), March 16 – 18, 2017, Chennai, India.

[11] ZouFutai, Gang Yuxiang, Pei Bei, ―Web Phishing Detection Based on Graph Mining‖ in 2016 2nd IEEE International Conference on Computer and Communications.

[12] Gang Liu, Bite Qiu, Liu Wenyin. ―Automatic Detection of Phishing Target from Phishing webpage[C]‖, 2010 20th International Conference on Pattern Recognition.Istanbul: IEEE Computer Society, 4153-4156, 2010.