A LITERATURE SURVEY ON

WEB PHISHING DETECTION

Domain: Applied Data Science

Team Id: PNT2022TMID29707

Batch no.: B11-5A1E

Team members: SHANMUGAPRIYA D (513119106074)

MAHALAKSHMI M (513119106050)

SINDHU S (513119106078)

SWETHA D (513119106089)

Paper 1: Real time detection of Phishing websites

Publication year: 2016

Authors: Abdulghani Ali Ahmed, Nurul Amirah Abdullah

Journal name: International Conference on Computer Applications & Information Security(ICCAIS)

Summary: Web Spoofing lures the user to interact with the fake websites rather than the real ones. The attacker creates a 'shadow' website that looks similar to the legitimate website. This paper proposes a detection technique of phishing websites based on checking Uniform Resources Locators(URLs) of web pages. A few selected features (if: IP address exist, URLs length > 54, Position of '//' >7, URLs contain '@', suffix or prefix '-') are used to differentiate between legitimate and spoofed web pages. In this experiment, Microsoft Visual Studio Express 2013 and C# language were used to create the application that differentiate them the designed application is named PhishChecker. When the user enters the URL, an alert pops up to indicate whether the website is secured or not. PhishChecker detect the phishing web pages with accuracy of 0.96. Moreover, the false negative rates does not exceed 0.105. This study only checks based on a few characteristics for detecting phishing attack.

Index terms: Uniform Resources Locator, PhishChecker, Microsoft Visual Studio Express 2013, C#

Paper 2: Detecting Phishing websites using Machine Learning

Publication year: 2019

Authors: Amani Alswailen, Bashayr Alabdullah, Norah Alrumayh, Dr. Aram Alsedrani

Journal name: International Conference and Workshop on Computing and Communication (IEMCON)

Summary: Phishing website is one of the internet security problems that target the human vulnerabilities rather than software vulnerabilities. All phishing website features (36 features) that have been used by previous researchers along with 3 new features have been used. Since the target of the phishing website is to steal sensitive information such as email or password, we consider the number of input that have the type of password as feature for phishing websites. We have noticed that a large number of phishing website doesn't use submit button instead they use a regular button, so they considered it as a feature. Features are extracted from URL, page content and page rank. To achieve high accuracy with minimal number of features, they deleted some features and achieved the same accuracy with 26 features that are used for Anti-phishing extension browser.

Index terms: Anti-phishing extension browser

Paper 3: Detection of Phishing websites by using machine learning-based URL analysis

Publication year: 2020

Authors: Mehmet Korkmaz, Ozgur Koray Sahingoz, Banu Diri

Journal name: International Conference on Computing, Communication and Networking Technologies(ICCNT)

Summary: Experienced attackers target on the weakness of the computer users by trying to phish them with bogus web pages. URLs have an important place in detecting phishing attacks especially for classifying the web page quickly. In this paper, a machine learning

based phishing detection was proposed using 8 different algorithms to analyze the URLs. The algorithms are Logistic Regression(LR), K-Nearest Neighbourhood(KNN), Support Vector Machine(SVM), Decision Tree(DT), Naive Bayes(NB), XGBoost, Random Forest(RF) and Artificial Neural Network(ANN). The model is trained by using Sklearn library in the Python programming language. Due to the structure of the trained system, the execution time for a single URL address and 100 URL has not a considerable difference. RF algorithm shows high accuracy rate. With the use of huge dataset, they plan to enhance their system by using some hybrid algorithms and also deep learning models.

Index terms: Sklearn library, Python, Hybrid algorithms, Deep learning

Paper 4: Detection of Phishing Websites from URLs by using Classification Techniques on WEKA

Publication Year: May 15,2021

Authors: Emre Kocyigit, Kubra Erensoy, Buket Geyik

Journal Name: International Conference on Innovative Computation Technologies[ICICT] IEEE

Summary: The Internet is getting stronger day by day and it makes our lives easier with many applications that are executed on the cyberworld. However, with the development of the internet, cyber-attacks have increased gradually and identity thefts have emerged. Traditional security mechanisms cannot prevent these attacks because they directly target the weakest part of connection end-users. Machine learning technology has been used to detect and prevent this type of intrusions. The anti-phishing method has been developed by detecting the attacks made with the technologies used. In this paper, they have executed a phishing detection system on WEKA and tested its efficiency by using a public dataset as Catch Phish D3 by using different classification techniques. To make this, it is needed to make some pre-processing steps to use the dataset in Weka system. Additionally, apart from the URL based features, some content-based features can also be used here. They also got help from some third-party organization/web pages as Alexa and Who is to identify whether the page is phishing or not.

Index terms: phishing attacks, machine learning, classification algorithms, phishing detection, cybersecurity

Paper 5: Machine Learning Techniques for Detection of Website Phishing: A Review for Promises and Challenge

Publication Year: May 17,2021

Authors: Ammar Odeh, Ismail Keshta, Eman Abdelfattah

Journal Name: International Conference on Innovative Computation Technologies [ICICT] IEEE

Summary: Websites phishing is a cyber-attack that targets online users to steal their sensitive information including login credentials and banking details. Several solutions to phishing websites attacks have been proposed such as heuristics, blacklist or whitelist, and Machine Learning (ML) based techniques. This survey paper also identifies deep learning-based techniques with better performance for detecting phishing websites compared to the conventional ML techniques. Machine learning (ML) is a multidisciplinary technique applied in supervised learning to construct predictive models. The problem of phishing attacks can be handled by transferring it to classification. Labelled historical data of websites is used to train and evaluate the model. By the integration of models into web browsers, phishing activities can be detected. Inefficiency of ML techniques on a large amount and images data, and websites with captcha information has been identified. Small size of datasets to train the ML techniques is another challenge as identified in this research. It is also suggested that an automated framework should be proposed based on ensemble learning and deep learning techniques in future works.

Index terms: ensemble learning, deep learning, cyber-attack

Paper 6: Phishing Website Detection using Machine Learning Algorithms

Publication Year: 2018

Authors: Rishikesh Mahajan, Irfan Siddarame

Journal Name: International Conference on Innovative Computation Technologies [ICICT] IEEE

Summary: Phishing attack is a simplest way to obtain sensitive information from innocent users. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This paper deals with machine learning technology for detection of phishing URLs by extracting and analysing various features of legitimate and phishing URLs. Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. Heuristic based detection which includes characteristics that are found to exist in phishing attacks in reality and can detect zero-hour phishing attack, but the characteristics are not guaranteed to always exist in such attacks and false positive rate in detection is very high This paper aims to enhance detection method to detect phishing websites using machine learning technology and achieved 97.14% detection accuracy using random forest algorithm with lowest false positive rate. Also result shows that classifiers give better performance when we used more data as training data.

Index terms: Phishing attack, Machine learning

Paper 7: A Research on Website Phishing Detection Based on LSTM RNN

Publication year: 2020

Author: Hochreiter and Schmidhuber

Journal Name: Information Technology, Networking,Electronics and Automation Control Conference.

Summary: Phishing attacks are growing threats to cyber security in worldwide. These criminals are usually profitable using phishing, so their goal usually is online banking, online payment platform, and mobile commerce applications. To make up for the shortcomings of blacklist technology, the use of machine learning algorithms to identify phishing links becomes the mainstream of current research. In order to effectively detect phishing attacks, this paper designed a new detection system for phishing websites using LSTM Recurrent Neural Networks (RNN). LSTM has the advantage of capturing data timing and longterm dependencies. LSTM has strong learning ability, can automatically learn data characterization without manual extraction of complex features, and has strong potential in the face of complex high-dimensional massive data. Experimental results

show that this model approach the accuracy of 99.1%, is higher than that of other neural network algorithms.

Index terms: LSTM Recurrent Neural Networks, machine learning

Paper 8: A Machine Learning Approach for URL Based Web Phishing Using Fuzzy Logic as Classifier

Publication year: 2019

Authors: Happy Chapla,Riddhi Kotak, Mittal Joiser

Journal Name: International Conference on Communication and Electronics Systems

Summary: Phishing is the major problem of the internet era. In this era of internet the security of our data in web is gaining an increasing importance. So the detection of phishing site is necessary. In website based phishing, website is copy of the original website which looks like same but, the aim is different. To overcome the problem of phishing we design a framework to detect it using the fuzzy logic as a classifier. For that we collect our dataset from the Phish tank site, Open phish site and the URL of the website which can live on the web. Our model constructed using both phishing and legitimate URLs including the features which have been extracted from these models. Fuzzy classifier is implemented using MatLab and the best results are achieved with 91.46% accuracy. Also the machine learning techniques play the important role as a classifier.

Index terms: fuzzy logic, MatLab, fuzzy classifier

Paper 9: A Methodical Overview on Detection, Identification and Proactive Prevention of Phishing Websites

Publication year: 2021

Authors: Bhagwat M. D,Dr. Patil P. H. , Dr. T. S. Vishawanath

Journal Name: International Conference on Intelligent Communication Technologies and Virtual Mobile Networks

Summary: Detecting and finding some phishing websites in real-time for a day now is really a dynamic. Fuzzy logic strategies may be an important method in detecting and testing phishing websites due to the ambiguities involved in the detection. Instead of exact principles, Fuzzy logic provides a more intuitive way of dealing with quality variables. An approach to fuzziness resolution and an open and intelligent phishing website detection model will be proposed in the Phishing website assessment. This approach is based on smooth logic and machine learning algorithms that define various factors on the phishing website. A total of 30 characteristics or features and phishing website attributes can be used for phishing detection with high accuracy. If the maximum numbers of features and phishing factors on the web site are considered as an input parameter for a fuzzy inference method and the rules basis for all the parameters can be set highly accurately, phishing can be detected. Rather than precise principles, Fuzzy provides a more natural way to deal with quality variables.

Index terms: fuzzy logic, phishing website attributes, machine learning

Paper 10: A Phishing Detection System based on Machine Learning

Publication year: 2019

Authors: Chee-Yu Wu,Cheng-Chung Kuo,Chu-Sing Yang

Journal Name: Institute of Computer and Communication Engineering National Cheng Kung University Tainan, Taiwan

Summary: Now a days Internet has become an essential part of human beings. Based on some weaknesses of human nature, hackers have designed confusing phishing pages to entice web viewers to proactively expose their privacy. Here, they proposed a URL-based detection system -combining the URL of the web page URL and the URL of the web page source code of the user. Due to this we implement Support-vector machine to be the machine learning algorithm model in our system. The system is designed to provide high accuracy and low false positive rate detection results for unknown phishing pages. In addition, the front-end part of the system will be designed to alert users when they access an phishing webpage via browsers. Through this the accuracy of the phishing detection system will be improved in order to against malicious from the dark side of the network.

Index terms: detection system, source code, machine learning, support-vector machine

Paper 11: Phishing Website Detection Using URL-Assisted

Publication Year: June 2014

Authors: Chon Lin Tan, Kang Leng Chiew, SaNah Sze

Journal Name: Faculty of Computer Science and Information Technology University Sarawak, Malaysia

Summary: They plan to design a more systematic data mining approach for parsing search engine results. They will also focus on automating the process of extracting ownership information from the WHOIS raw data. It would also be interesting to experiment with different search engines to evaluate their performance for our anti-phishing solution. Another direction is to incorporate the brand logo into the proposed method to offer a greater contribution in website identity discovery. Brand names are as legitimate websites. They exploited this Phishing pattern and proposed a URL-assisted brand name Weighting system to detect phishing websites. The webpage textual content to be fed to search engine. From the search results, we obtain the Domain name that occurs the most number of times, followed by looking up the domain name owner in WHOIS database. A query website is considered as phishing when its domain Name owner does not match the domain name owner from the Legitimate website.

Index terms: brand names, WHOIS database, data mining

Paper 12: Phishing Website Detection Framework Through Web Scraping and Data Mining

Publication Year: 2017

Authors: Andrew J. Park, Ruhi Naaz Quadari, Herbert H. Tsang

Journal Name: Department of Computing Science Thompson Rivers University Kamloops, BC, Canada.

Summary: This paper has presented a prototype work of a phishing website detection framework, it is based on content-based heuristics. The heuristics were generated with training data sets collected from active phishing websites and collections of previously de-tested phishing websites. The crawler scraped relevant information from the data sets. Rapid Miner was used to the scraped information and relevant heuristics were identified. These heuristics were further analysed. The dynamic aspect of heuristic and phishing factors were assumed since attackers construct intelligently to avoid all phishing

detection tools. The future research plan that will further develop Phishing-Detective includes: Discovering and plug-in/extension for web browsers that is analyzing more contributing heuristics in detecting phishing websites; Automating the calculation of heuristics' weights and discovery of new heuristics dynamically using big data, data mining, and machine learning techniques; and Developing a embedded so that it detects phishing websites in real time while users browse web.

Index terms: content based heuristics, rapid miner