Web Phishing Detection - Literature Review

COMPUTER SCIENCE AND ENGINEERING
2022 - SEPTEMBER

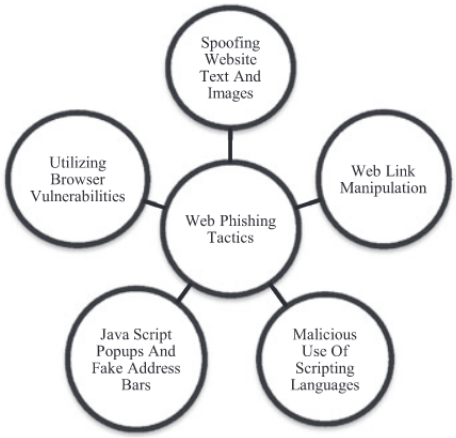
Contributors:

A Kirthic Vishnu	:	19z301@psgtech.ac.in
Adharsh S	:	19z302@psgtech.ac.in
Kumaresh S	:	19z327@psgtech.ac.in
Mridula M	:	19z332@psgtech.ac.in

Research Paper - 1

A survey and classification of web phishing detection schemes[1]

+ The paper highlights the type of attacks possible taking various vectors into consideration and classifies them as shown in the figure below



+ It also highlights various phishing preventive solutions based on watermarking, RFID technology, QR code based,etc. The paper also mentions various detection schemes and highlights their pros and cons like:
    + ***Search engine-based techniques:***
      These techniques collect text, photos, or URLs from websites and utilize them as search terms to gauge how popular a website is in order to identify phishing. The methods differ, though, in terms of I the type and quantity of features extracted from a webpage (text, URL, or images), (ii) the number of search engines used to gauge the popularity of the webpage, (iii) the number of top results used for matching,
    + ***Heuristics and machine learning-based techniques:***
      This group of techniques extracts a number of features, such as web page content, URLs, and/or network information, and a set of machine learning or classification algorithms, which are then used to build a model for classification.
+ The study offers an analysis of the methods suggested for phishing detection that has been carried out. The study focuses on the fact that phishing detection methods outperform phishing prevention and user education strategies since they don't call for modifications to authentication systems and don't rely on the user's capacity to recognize phishing. Regarding the additional hardware needed and password management, phishing detection solutions are additionally less expensive than phishing prevention solutions.
+ Because they just need a single search engine query result and its underlying algorithm to identify phishing websites at the user's end, search engine-based strategies have been shown to be the lightest solutions for phishing detection.

Research Paper - 2

Intelligent web-phishing detection and protection scheme using integrated features of Images, frames, and text[2]

The combined features of the text, graphics, and frames are used in this paper's robust adaptive neuro-fuzzy inference system (ANFIS)-based web-phishing detection and protection approach. The integrated elements of phishing websites' graphics, frames, and text were used in this work to propose an intelligent phishing detection and defense strategy. Based on the plans outlined in Aburrous et al. (2010) and Barraclough et al. (2014), an effective ANFIS algorithm was created, examined, and verified for phishing website identification and protection (2015). The study employs a hybrid technique to choose text features, utilizing factors such as Page rank, Google Index, long URLs, domain entity relationships, and many others.

Research Paper - 3

Intelligent cyber-phishing detection for online[3]

+ The approach is based on multiple ML classification algorithms, using ANFIS implemented in MATLAB toolbox with features and NB, PART, J48, JRip implemented in Waikato Environment Knowledge Analysis (WEKA).
+ Methodology combines blacklist-based features, web content-based and heuristic-based approaches enabling a set of data (phishing websites, suspicious websites, spoofed web and legitimate websites) to be extracted from diverse sources.
+ Based on evaluation of proposed methodology using ANFIS and MATLAB, using randomised and time-based evaluation, the method achieved 98.1% accuracy with 1.9% average testing error. Upon fine tuning, 99% accuracy was achieved with a 0.1% average error. Time taken to build the model was 0.01 Secs. The results demonstrate that the method can generalise well to new phishing attack
+ J48 classifier shows a good performance that is 99.3% instances are correctly classified (TP), while 0.66% instances are incorrectly classified. Time taken to build the model is 0.01 secs
+ NB classifier, on applying Randomised and Time-based evaluation methods, has high performance. 99.3% phishing instances correctly classified (TP) and 0.66% instances are incorrectly classified, while Time taken to build the model was 0.01 Secs. Recall maintained the accuracy, while F-Measure average score slightly decreased by 0.03%. ROC curve average score decreased further by 0.5%, Precision score was the lowest with 98.7% accuracy. The overall accuracy (99.3%) exceeds 95%
+ PART classifier randomly split the features equally into Tens, trained on training-sets and test on unseen test-set. t PART algorithm achieved high performance, obtaining 99.3% accuracy (correctly classified as TP) and 0.66% instances incorrectly classified. Time taken to build the model was 0.01 to 0.006 Secs, which was the best speed

+ JRip demonstrated best performance, obtaining 99.3% accuracy with speed of 0.07 to 0.14 Secs with 0.66% instances incorrectly classified. This is the worst achievement in speed
+ The proposed novel methodology identifies abnormal features in URLs, links and images, text, forms, frames links and files. The results show that the approach can also classify between phishing, suspicious and legitimate websites accurately
+ The method with 0.6 error rates could be improved by deploying constant flow of emerging features specifically dedicated to detecting fraudulent websites. Whilst this method has produced excellent results, it is important to have continuous features update in order to keep ahead of evolving phishing strategies (X

Research Paper - 4

Boosting the Accuracy of Phishing Detection with Less Features Using XGBOOST[4]

+ Use of Anti-Phishing and Network analysis tool (CANTINA) heuristic features with a new additional attributes that are combined with the new ones (Domain Top page Similarity) to detect Phishing pages
+ Definition of a set of rules in order to identify Phishing web pages, Two groups of rules were distinguished: -the simple rules, based on web page URL and the higher complex and time consuming rules based on the analysis of metadata, query search engines and blacklists the search engine based rules, the red aged keywords based rules, the obfuscation based rules, the blacklists based rules, the reputation based rules, the content based rules.
+ Extraction of identities from some features such as Meta title, meta description, content attributes and "href" attributes of tag for detecting Phishing attacks.
+ Use of feature vector of size 23, in which four were structural features from URLs, nine were lexical features and ten were features targeting mostly brand and websites and classification using SVM.
+ Evaluation of feature selection techniques : The correlation based and wrapper based feature selection techniques.

+ Detection of hidden URLs using lexical features
+ Use of lexical and domain features extracted from Phishing URLs to detect Phishing websites
+ Use of Tabsol to fight against Tab nabbing(Tab nabbing is a variant of Phishing attack in which a malicious page opened in a tab disguises itself to a popular website's login page such as Gmail, Facebook login pages with the objective to steal credentials)
+ Comparison of different features assessment techniques in the website phishing context in order to determine the minimal set of features for detecting phishing activities

1. Experiment shows that the comparison of the accuracy of algorithms for Different Feature Groups based on the decisive values of the features demonstrated that best accuracy is obtained for Random Forest by 96.07% , but random forests have been observed to overfit some datasets and noisy classification tasks.
2. Use of Probabilistic Neural Networks(PNNs) and e integration of PNN with K-medoids clustering to significantly reduce complexity without jeopardizing the detection accuracy. The experimental results show that 96.79% accuracy is achieved with low false errors.
3. Use of XGBOOST algorithm which improved the performance that a predictive model can achieve in the task of phishing website detection
4. Comparison of XGBOOST with Probabilistic Neural Networks (PNN) and Random forest (RF) methodin which all the methods (classifiers) were trained and tested using the same dataset and evaluated using the same performance metrics for a fair comparison, XGBOOST algorithm returned the accuracy of 97.27%

Research Paper 5

A Deep Learning Technique for Web Phishing Detection Combined URL Features and Visual Similarity[5]

+ Listed ways in which phishing detection is usually performed
    1. DOM (Document Object Model) tree analysis
    2. Visual feature based technique
    3. CSS(Cascading Style Sheets) based similarity analysis
    4. Website image comparison
    5. Visual perception method
    6. Hybrid Method
+ Use of Convolutional Neural Networks(CNNs) to detect web phishing attacks based on URLs and screenshots of websites , as CNN is best in processing high dimensional data such as videos and images
+ Division of snapshots of legitimate and suspected pages into blocks and matching using Earth Mover's algorithm , which gives high detection rate(99.6%)
+ Normalized Compression Distance (NCD) to compute similarities based on distance between the image of a requested website and the image of a cached benign website , which gives a high true positive rate but is practically impossible with real time browsing
+ Formulation of web phishing detection as a binary classification problem with the URL and images of websites as input leading to their classification as either legitimate websites or phished websites which can detect newly created phishing webpages based only on the URL and the screenshot of suspicious websites. The proposed model shows a classification accuracy of 99.67%

References

1. Varshney, G., Misra, M., and Atrey, P. K. (2016) A survey and classification of web phishing detection schemes. *Security Comm. Networks*, 9: 6266– 6284. doi: 10.1002/sec.1674.
2. M.A. Adebowale, K.T. Lwin, E. Sánchez, M.A. Hossain, Intelligent web-phishing detection and protection scheme using integrated features of Images, frames and text, Expert Systems with Applications, Volume 115, 2019, Pages 300-313, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2018.07.067.
3. Barraclough, P. A., Fehringer, G., & Woodward, J. (2021). Intelligent cyber-phishing detection for online. Computers & Security, 104, 102123. https://doi:10.1016/j.cose.2020.102123
4. Hajara, Musa & Gital, A & Mohzo, & Bitrus, Gideon & Jumaat, Nurul & Juma'at, & Muhammad, & Balde, Abubakar. (2020). Boosting the accuracy of phishing detections with less features using XGBOOST. 8. 81-90. https://www.researchgate.net/publication/339676208_Boosting_the_accuracy_of_phishing_detections_with_less_features_using_XGBOOST
5. Al-Ahmadi, Saad, A Deep Learning Technique for Web Phishing Detection Combined URL Features and Visual Similarity (2020). International Journal of Computer Networks & Communications (IJCNC) Vol.12, No.5, September 2020, Available at SSRN: https://ssrn.com/abstract=3716033