APPLIED DATA SCIENCE

WEB PHISHING DETECTION

LITERATURE SURVEY

BATCH: B2-2M4E

TEAM LEADER:

ARIVAZHAGAN.R

TEAM MEMBERS:

DHANISHSRIRAM.K

DHIVAKAR.K

DINESH.K

WEB PHISHING DETECTION

LITERRATURE SURVEY

Phishing is a form of fraud in which the attacker tries to learn sensitive information such as login credentials or account information by sending as a reputable entity or person in email or other communication channels. Typically, a victim receives a message that appears to have been sent by a known contact or organization. The message contains malicious software targeting the user's computer or has links to direct victims to malicious websites in order to trick them into divulging personal and financial information, such as passwords, account IDs or credit card details. Phishing is popular among attackers, since it is easier to trick someone into clicking a malicious link which seems legitimate than trying to break through a computer's defense systems.

1. TITLE **:** Anomaly Based Web Phishing Page Detection

AUTHORS **:** Ying Pan, Xuhua Ding

PUBLISHED IN **:** 2006

Many anti-phishing schemes have recently been proposed in literature. Despite all those efforts, the threat of phishing attacks is not mitigated. One of the main reasons is that phishing attackers have the adaptability to change their tactics with little cost. In this paper, we propose a novel approach, which is independent of any specific phishing implementation. Our idea is to examine the anomalies in Web pages, in particular, the discrepancy between a Web site's identity and its structural features and HTTP transactions. It demands neither user expertise nor prior knowledge of the Web site. The evasion of our phishing detection entails high cost to the adversary. As shown by the experiments, our phishing detector functions with low miss rate and low false positive rate.

2. TITLE **:** Phishing Web Page Detection with HTML

AUTHORS **:** Linshu Ouyang, Yongzheng Zhang

PUBLISHED IN **:** 2021

Phishing web page is one of the most serious threats to the users of the Internet. Traditional phishing web page detection methods rely on manually designed features. Recently, deep learning-based methods using HTML as input have achieved significant detection performance improvement. They usually treat HTML codes as sequences of characters and utilize Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN) for classification. However, CNN and RNN typically can only extract local features in the HTML code sequences while failing to model the long-range semantics that is crucial for phishing detection. In this paper, we propose a novel Graph Neural Network (GNN) based phishing web page detection method that can effectively utilize the inherent structural information of HTML to capture the long range semantics. We first naturally represent an HTML as a graph according to its Document Object Model (DOM) and utilize RNN to extract the local features of node attributes. Then we adopt GNN to model the long-range relations between nodes based on these local features and the graph structure. Our proposed model combines the advantage of RNN and GNN to better understand the intention of HTML codes. Extensive experiments on a real-world dataset demonstrate that the accuracy of our method outperforms other state-of-the-art methods by a large margin.

3. TITLE **:** Phishing Web Page Detection Using Optimized Machine Learning

AUTHORS **:** Jordan Stobbs, Biju Issac, Seibu Mary Jacob

PUBLISHED IN **:** 2020

Phishing is a type of social engineering attack that can affect any company or anyone. This paper explores the effect that different features and optimisation techniques have on the accuracy of intelligent phishing detection using machine learning algorithms. This work looks at both hyperparameter optimisation as well as

feature selection optimisation. For hyperparameter tuning, both TPE (Tree-structured Parzen Estimator) and GA (Genetic Algorithm) were tested, with the best option being model dependent.

For feature selection, GA, MFO (Moth Flame Optimisation) and PSO (Particle Swarm Optimisation) were used with PSO working best with a Random Forest model. This work used URL (Uniform Resource Locator), DOM (Document Object Model) structure, page rank and page information related features. This research found that the best combination was Random Forest using PSO for feature selection and TPE for hyperparameter optimisation, giving an accuracy of 99.33%.

4. TITLE **:** An Improved Genetic Algorithm for Web Phishing Detection Feature Selection

AUTHORS **:** Jiachen Wang

PUBLISHED IN **:** 2022

Feature selection is a useful dimension reduction method in data preprocessing for machine learning. It is an discrete optimization problem in essence. Genetic algorithm is a meta heuristic optimization algorithm for discrete optimization problem, which simulates the biological evolution behavior in nature. The original Genetic algorithm, like many other swarm intelligence optimization algorithms, has its own defects, easy to fall into local optimum, and slow convergence speed. To improve the performance of Genetic algorithm, an improved algorithm ECGA is proposed. In order to measure the diversity of population, the information entropy of population is calculated after fitness calculation. To speed up convergence, the consensus mechanism is presented after mutation which is used as a alternate mutation operator. And finally, ECGA is applied to feature selection to test its actual effect. The experimental results show that the improved Genetic algorithm is better than the feature selection algorithm proposed in recent 2 years for the web phishing detection problem. Our algorithm achieves a average F1socre of 97%.

5. TITLE : User-Centric Phishing Threat Detection

AUTHORS : Yuen-Hsien Tseng

PUBLISHED IN : 2020

This paper presents a context-aware phishing threat detection model from users' behavioral perspectives. The context of users' information accesses is investigated to explore the users' browsing behaviors that confront phishing situations. Large-scale experiments show that our approach achieves an accuracy of 0.9973 and an F1 score of 0.9311 for predicting the phishing threats of users' next accesses without intelligent content analysis. Error analysis indicates that our proposed model results in a favorably low false positive rate of 0.0006. In practice, our proposed model is complementary to the existing anti-phishing techniques for cost-effectively blocking phishing threats with wisdom of the crowds.

6. TITLE : Using Domain Top-page Similarity Feature in Machine Learning Based Web Phishing Detection

AUTHORS : Nuttapong Sanglerdsinlapachai, Arnon Rungsawang

PUBLISHED IN : 2010

This paper presents a study on using a concept feature to detect web phishing problem.

Following the features introduced in Carnegie Mellon Anti-phishing and Network Analysis Tool (CANTINA), we applied additional domain top-page similarity feature to a machine learning based phishing detection system. We preliminarily experimented with a small set of 200 web data, consisting of 100 phishing webs and another 100 non phishing webs. The evaluation result in terms of f-measure was up to 0.9250, with 7.50% of error rate.

7. TITLE **:** Detecting Phishing Web Pages with Visual Similarity Assessment Based on Earth Mover's Distance (EMD)

AUTHORS **:** Anthony Y. Fu, Liu Wenyin, Xiaotie Deng

PUBLISHED IN **:** 2006

An effective approach to phishing Web page detection is proposed, which uses Earth Mover's Distance (EMD) to measure Web page visual similarity. We first convert the involved Web pages into low resolution images and then use color and coordinate features to represent the image signatures. We use EMD to calculate the signature distances of the images of the Web pages. We train an EMD threshold vector for classifying a Web page as a phishing or a normal one. Large-scale experiments with 10,281 suspected Web pages are carried out to show high classification precision, phishing recall, and applicable time performance for online enterprise solution. We also compare our method with two others to manifest its advantage. We also built up a real system which is already used online and it has caught many real phishing cases.

8. TITLE **:** What is the Cyber Kill Chain | IEEE Computer Society

AUTHORS **:** Pratik Dholakiya

PUBLISHED IN **:** 2018

The cyber kill chain is essentially a cybersecurity model created by Lockheed Martin that traces the stages of a cyber-attack, identifies vulnerabilities, and helps security teams to stop the attacks at every stage of the chain. The term kill chain is adopted from the military, which uses this term related to the structure of an attack. It consists of identifying a target, dispatch, decision, order, and finally, destruction of the target.