A survey and classification of web phishing detection schemes

Authors: Gaurav Varshney, Manoj Misra, Pradeep K.Atrey
Published on: 26 October 2016.

Abstract:

Phishing is a fraudulent technique that is used over the Internet to deceive users with the goal of extracting their personal information such as username, passwords, credit card, and bank account information. The key to phishing is deception. Phishing uses email spoofing as its initial medium for deceptive communication followed by spoofed websites to obtain the needed information from the victims. Phishing was discovered in 1996, and today, it is one of the most severe cybercrimes faced by the Internet users. Researchers are working on the prevention, detection, and education of phishing attacks, but to date, there is no complete and accurate solution for thwarting them. This paper studies, analyzes, and classifies the most significant and novel strategies proposed in the area of phished website detection, and outlines their advantages and drawbacks. Furthermore, a detailed analysis of the latest schemes proposed by researchers in various subcategories is provided. The paper identifies advantages, drawbacks, and research gaps in the area of phishing website detection that can be worked upon in future research and developments. The analysis given in this paper will help academia and industries to identify the best anti-phishing technique.

Web Phishing Detection Using a Deep Learning Framework

Authors: Ping Yi, Yuxiang Guan, Futai Zou, Yao Yao, Wei Wang, Ting Zhu.
Published on: 26 Sept 2018

Abstract:

Web service is one of the key communications software services for the Internet. Web phishing is one of many security threats to web services on the Internet. Web phishing aims to steal private information, such as usernames, passwords, and credit card details, by way of impersonating a legitimate entity. It will lead to information disclosure and property damage. This paper mainly focuses on applying a deep learning framework to detect phishing websites. This paper first designs two types of features for web phishing: original features and interaction features. A detection model based on Deep Belief Networks (DBN) is then presented. The test using real IP flows from ISP (Internet Service Provider) shows that the detecting model based on DBN can achieve an approximately 90% true positive rate and 0.6% false positive rate.

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

Authors: M.A. Adebowale, K.T. Lwin, E. Sánchez, M.A. Hossain
Published on: 25 April 2018

Abstract:

A phishing attack is one of the most significant problems faced by online users because of its enormous effect on the online activities performed. In recent years, phishing attacks continue to escalate in frequency, severity and impact. Several solutions, using various methodologies, have been proposed in the literature to counter the web-phishing threats. Notwithstanding, the existing technology cannot detect the new phishing attacks accurately due to the insufficient integration of features of the text, image and frame in the evaluation process. The use of related features of images, frames and text of legitimate and non-legitimate websites and associated artificial intelligence algorithms to develop an integrated method to address these together. This paper presents an Adaptive Neuro-Fuzzy Inference System (ANFIS) based robust scheme using the integrated features of the text, images and frames for web-phishing detection and protection. The proposed solution achieves 98.3% accuracies. To our best knowledge, this is the first work that considers the best-integrated text, image and frame feature based solution for phishing detection scheme.

Systematization of Knowledge (SoK): A Systematic Review of Software-Based Web Phishing Detection

Authors: Zuochao Dou; Issa Khalil; Abdallah Khreishah; Ala Al-Fuqaha; Mohsen Guizani
Published on: **13 September 2017**

Abstract:

Phishing is a form of cyber attack that leverages social engineering approaches and other sophisticated techniques to harvest personal information from users of websites. The average annual growth rate of the number of unique phishing websites detected by the Anti Phishing Working Group is 36.29% for the past six years and 97.36% for the past two years. In the wake of this rise, alleviating phishing attacks has received a growing interest from the cyber security community. Extensive research and development have been conducted to detect phishing attempts based on their unique content, network, and URL characteristics. Existing approaches differ significantly in terms of intuitions, data analysis methods, as well as evaluation methodologies. This warrants a careful systematization so that the advantages and limitations of each approach, as well as the applicability in different contexts, could be

analyzed and contrasted in a rigorous and principled way. This paper presents a systematic study of phishing detection schemes, especially software based ones. Starting from the phishing detection taxonomy, we study evaluation datasets, detection features, detection techniques, and evaluation metrics. Finally, we provide insights that we believe will help guide the development of more effective and efficient phishing detection schemes.

Web phishing detection techniques: a survey on the state-of-the-art, taxonomy and future directions

Authors: M. Vijayalakshmi,S. Mercy Shalinie,Ming Hour Yang,Raja Meenakshi U.
First published: 23 September 2020

Abstraction:

Internet dragged more than half of the world's population into the cyber world. Unfortunately, with the increase in internet transactions, cybercrimes also increase rapidly. With the anonymous structure of the internet, attackers attempt to deceive the end-users through different forms namely phishing, malware, SQL injection, man-in-the-middle, domain name system tunnelling, ransomware, web trojan, and so on. Amongst them, phishing is the most deceiving attack, which exploits the vulnerabilities in the end-users. Phishing is often done through emails and malicious websites to lure the user by posing themselves as a trusted entity. Security experts have been proposing many anti-phishing techniques. Till today there is no single solution that is capable of mitigating all the vulnerabilities. A systematic review of current trends in web phishing detection techniques is carried out and a taxonomy of automated web phishing detection is presented. The objective of this study is to acknowledge the status of current research in automated web phishing detection and evaluate their performance. This study also discusses the research avenues for future investigation.

Using Domain Top-page Similarity Feature in Machine Learning-Based Web Phishing Detection

Authors: Nuttapong Sanglerdsinlapachai; Arnon Rungsawang
Published on: 2010 Third International Conference on Knowledge Discovery and Data Mining.

Abstract:
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.

Web Phishing Detection Based on Page Spatial Layout Similarity

Authors: Weifeng, Hua Lu, Baowen Xu, Hongji Yang
Published On: July 8, 2012

Abstract:
Web phishing is becoming an increasingly severe security threat in the web domain. Effective and efficient phishing detection is very important for protecting web users from loss of sensitive private information and even personal properties. One of the keys of phishing detection is to efficiently search the legitimate web page library and to find those page that are the most similar to a suspicious phishing page. Most existing phishing detection methods are focused on text and/or image features and have paid very limited attention to spatial layout characteristics of web pages. In this paper, we propose a novel phishing detection method that makes use of the informative spatial layout characteristics of web pages. In particular, we develop two different options to extract the spatial layout features as rectangle blocks from a given web page. Given two web pages, with their respective spatial layout features, we propose a page similarity definition that takes into account their spatial layout characteristics. Furthermore, we build an R-tree to index all the spatial layout features of a legitimate page library. As a result, phishing detection based on the spatial layout feature similarity is facilitated by relevant spatial queries via the R-tree. A series of simulation experiments are conducted to evaluate our proposals. The results demonstrate that the proposed novel phishing detection method is effective and efficient.

Anomaly Based Web Phishing Page Detection

Authors: Ying Pan; Xuhua Ding
Published On: 2006 22nd Annual Computer Security Applications Conference (ACSAC'06)

Abstract:
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.

PhishZoo: An Automated Web Phishing Detection Approach Based on Profiling and Fuzzy Matching

Authors: Sadia Afroz and Rachel Greenstadt

Abstract:
Phishing is a web-based attack that uses social engineering techniques to exploit Internet users and acquire sensitive data. Most phishing attacks work by creating a fake version of the real site’s web interface to gain the user’s trust. Despite the fact that these phishing sites look identical or nearly identical to the real sites they imitate, user studies have shown that users ignore browser-based indicators and often use the appearance of a site to judge the authenticity of sites, just as they use the appearance of physical sites to judge their authenticity. This paper proposes a phishing detection approach—PhishZoo—that uses profiles of trusted websites’ appearances built with fuzzy hashing techniques to detect phishing. We evaluate our approach on over 600 phishing sites imitating 20 real sites and show that it provides similar accuracy to blacklisting approaches, with the advantage that it can classify new attacks and targeted attacks against smaller sites (such as corporate intranets). PhishZoo has the potential to have a beneficial impact on the phishing “arms race”by reducing the effectiveness of sites that look too much like the real sites and thus giving users a chance to detect sites that “look phishy.”

Detecting visually similar Web pages: Application to phishing detection

Authors: Teh-Chen, Scott Dick, James Miller.
Published on: 10 June 2010

Abstract:

We propose a novel approach for detecting visual similarity between two Web pages. The proposed approach applies Gestalt theory and considers a Web page as a single indivisible entity. The concept of supersignals, as a realization of Gestalt principles, supports our contention that Web pages must be treated as indivisible entities. We objectify, and directly compare, these indivisible supersignals using algorithmic complexity theory. We illustrate our approach by applying it to the problem of detecting phishing scams. Via a large-scale, real-world case study, we demonstrate that 1) our approach effectively detects similar Web pages; and 2) it accuractely distinguishes legitimate and phishing pages.

Machine learning based phishing detection from URLs

Authors: Ozgur KoraySahingoz[a]EbubekirBuber[b]OnderDemir[b]BanuDiri[c].

Published on: 7 May 2018

Abstract:

Due to the rapid growth of the Internet, users change their preference from traditional shopping to the electronic commerce. Instead of bank/shop robbery, nowadays, criminals try to find their victims in the cyberspace with some specific tricks. By using the anonymous structure of the Internet, attackers set out new techniques, such as phishing, to deceive victims with the use of false websites to collect their sensitive information such as account IDs, usernames, passwords, etc. Understanding whether a web page is legitimate or phishing is a very challenging problem, due to its semantics-based attack structure, which mainly exploits the computer users' vulnerabilities. Although software companies launch new anti-phishing products, which use blacklists, heuristics, visual and machine learning-based approaches, these products cannot prevent all of the phishing attacks. In this paper, a real-time anti-phishing system, which uses seven different classification algorithms and natural language processing (NLP) based features, is proposed. The system has the following distinguishing properties from other studies in the literature: language independence, use of a huge size of phishing and legitimate data, real-time execution, detection of new websites, independence from third-party services and use of feature-rich classifiers. For measuring the performance of the system, a new dataset is constructed, and the experimental results are tested on it. According to the experimental and comparative results from the implemented

classification algorithms, Random Forest algorithm with only NLP based features gives the best performance with the 97.98% accuracy rate for detection of phishing URLs.

Phishing Detection Using Machine Learning Techniques

Authors: Vahid Shahrivari, Mohammad Mahdi Darabi, Mohammad Izadi

Published on: Sun, 20 Sep 2020

Abstract:

The Internet has become an indispensable part of our life, However, It also has provided opportunities to anonymously perform malicious activities like Phishing. Phishers try to deceive their victims by social engineering or creating mock-up websites to steal information such as account ID, username, password from individuals and organizations. Although many methods have been proposed to detect phishing websites, Phishers have evolved their methods to escape from these detection methods. One of the most successful methods for detecting these malicious activities is Machine Learning. This is because most Phishing attacks have some common characteristics which can be identified by machine learning methods. In this paper, we compared the results of multiple machine learning methods for predicting phishing websites.

Survey of review spam detection using machine learning techniques

Authors: Taghi M. Khoshgoftaar, Joseph D. Prusa, Aaron N. Richter & Hamzah Al Najada

Abstract:

Online reviews are often the primary factor in a customer's decision to purchase a product or service, and are a valuable source of information that can be used to determine public opinion on these products or services. Because of their impact, manufacturers and retailers are highly concerned with customer feedback and reviews. Reliance on online reviews gives rise to the potential concern that wrongdoers may create false reviews to artificially promote or devalue products and services. This practice is known as Opinion (Review) Spam, where spammers manipulate and poison reviews (i.e., making fake, untruthful, or deceptive reviews) for profit or gain. Since not all online reviews are truthful and trustworthy, it is important to develop techniques for detecting review spam. By extracting meaningful features from the text using Natural Language Processing (NLP), it is possible to conduct review spam detection using various machine learning techniques. Additionally, reviewer

information, apart from the text itself, can be used to aid in this process. In this paper, we survey the prominent machine learning techniques that have been proposed to solve the problem of review spam detection and the performance of different approaches for classification and detection of review spam. The majority of current research has focused on supervised learning methods, which require labeled data, a scarcity when it comes to online review spam. Research on methods for Big Data are of interest, since there are millions of online reviews, with many more being generated daily. To date, we have not found any papers that study the effects of Big Data analytics for review spam detection. The primary goal of this paper is to provide a strong and comprehensive comparative study of current research on detecting review spam using various machine learning techniques and to devise methodology for conducting further investigation.

Phishing Detection Based on Machine Learning and Feature Selection Methods

Authors: Mohammad Almseidin, AlMaha Abu Zuraiq, Mouhammd Al-kasassbeh, Nidal Alnidami,

Abstract:

With increasing technology developments, the Internet has become everywhere and accessible by everyone. There are a considerable number of web-pages with different benefits. Despite this enormous number, not all of these sites are legitimate. There are so-called phishing sites that deceive users into serving their interests. This paper dealt with this problem using machine learning algorithms in addition to employing a novel dataset that related to phishing detection, which contains 5000 legitimate web-pages and 5000 phishing ones. In order to obtain the best results, various machine learning algorithms were tested. Then J48, Random forest, and Multilayer perceptron were chosen. Different feature selection tools were employed to the dataset in order to improve the efficiency of the models. The best result of the experiment achieved by utilizing 20 features out of 48 features and applying it to Random forest algorithm. The accuracy was 98.11%

A machine learning based approach for phishing detection using hyperlinks information

Authors: Ankit Kumar Jain, B. B. Gupta

Abstract:

This paper presents a novel approach that can detect phishing attack by analysing the hyperlinks found in the HTML source code of the website. The proposed approach incorporates various new outstanding hyperlink specific features to detect phishing attack. The proposed approach has divided the hyperlink specific features into 12 different categories and used these features to train the machine learning algorithms. We have evaluated the performance of our proposed phishing detection approach on various classification algorithms using the phishing and non-phishing websites dataset. The proposed approach is an entirely client-side solution, and does not require any services from the third party. Moreover, the proposed approach is language independent and it can detect the website written in any textual language. Compared to other methods, the proposed approach has relatively high accuracy in detection of phishing websites as it achieved more than 98.4% accuracy on logistic regression classifier.

Phishing Detection Using Machine Learning Technique

Author: Junaid Rashid; Toqeer Mahmood; Muhammad Wasif Nisar; Tahira Nazir
Published in: 2020 First International Conference of Smart Systems and Emerging Technologies

Abstract:

Today, everyone is highly dependent on the internet. Everyone performed online shopping and online activities such as online Bank, online booking, online recharge and more on internet. Phishing is a type of website threat and phishing is Illegally on the original website Information such as login id, password and information of credit card. This paper proposed an efficient machine learning based phishing detection technique. Overall, experimental results show that the proposed technique, when integrated with the Support vector machine classifier, has the best performance of accurately distinguishing 95.66% of phishing and appropriate websites using only 22.5% of the innovative functionality. The proposed technique exhibits optimistic results when benchmarking with a range of standard phishing datasets of the “University of California Irvine (UCI)” archive. Therefore, proposed technique is preferred and used for phishing detection based on machine learning.

Phishing detection: A recent intelligent machine learning comparison based on models content and features

Authors: Neda Abdelhamid; Fadi Thabtah; Hussein Abdel-jaber
Published in: 2017 IEEE International Conference on Intelligence and Security Informatics (ISI)

Abstract:
In the last decade, numerous fake websites have been developed on the World Wide Web to mimic trusted websites, with the aim of stealing financial assets from users and organizations. This form of online attack is called phishing, and it has cost the online community and the various stakeholders hundreds of million Dollars. Therefore, effective counter measures that can accurately detect phishing are needed. Machine learning (ML) is a popular tool for data analysis and recently has shown promising results in combating phishing when contrasted with classic anti-phishing approaches, including awareness workshops, visualization and legal solutions. This article investigates ML techniques applicability to detect phishing attacks and describes their pros and cons. In particular, different types of ML techniques have been investigated to reveal the suitable options that can serve as anti-phishing tools. More importantly, we experimentally compare large numbers of ML techniques on real phishing datasets and with respect to different metrics. The purpose of the comparison is to reveal the advantages and disadvantages of ML predictive models and to show their actual performance when it comes to phishing attacks. The experimental results show that Covering approach models are more appropriate as anti-phishing solutions, especially for novice users, because of their simple yet effective knowledge bases in addition to their good phishing detection rate.

Improved Phishing Detection using Model-Based Features

Authors: Andr´e Bergholz, Gerhard Paaß, Frank Reichartz, Siehyun Strobel

Abstract:
Phishing emails are a real threat to internet communication and web economy. Criminals are trying to convince unsuspecting online users to reveal passwords, account numbers, social security numbers or other personal information. Filtering approaches using blacklists are not completely effective as about every minute a new phishing scam is created. We investigate the statistical filtering of phishing emails, where a classifier is trained on characteristic features of existing emails and subsequently is able to identify new phishing emails with different contents. We propose advanced email features generated by adaptively trained Dynamic Markov Chains and by novel latent Class-Topic Models. On a publicly

available test corpus classifiers using these features are able to reduce the number of misclassified emails by two thirds compared to previous work. Using a recently proposed more expressive evaluation method we show that these results are statistically significant. In addition we successfully tested our approach on a non-public email corpus with a real-life composition.

PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning

Authors: Ankit Kumar Jain, B. B. Gupta

Abstract:

Today, phishing is one of the most serious cyber-security threat in which attackers steal sensitive information such as personal identification number (PIN), credit card details, login, password, etc., from Internet users. In this paper, we proposed a machine learning based anti-phishing system (i.e., named as PHISH-SAFE) based on Uniform Resource Locator (URL) features. To evaluate the performance of our proposed system, we have taken 14 features from URL to detect a website as a phishing or non-phishing. The proposed system is trained using more than 33,000 phishing and legitimate URLs with SVM and Naïve Bayes classifiers. Our experiment results show more than 90% accuracy in detecting phishing websites using SVM classifier.