WEB PHISHING DETECTION

INTRODUCTION

Web service is a communication protocol and software between two electronic devices over the Internet [1]. Web services extends the World Wide web infrastructure to provide the methods for an electronic device to connect to other electronic devices [2]. Web services are built on top of open communication protocols such as TCP/IP, HTTP, Java, HTML, and XML. Web service is one of the greatest inventions of mankind so far, and it is also the most profound manifestation of computer influence on human beings [3].

With the rapid development of the Internet and the increasing popularity of electronic payment in web service, Internet fraud and web security have gradually been the main concern of the public [4]. Web Phishing is a way of such fraud, which uses social engineering technique through short messages, emails, and WeChat [5] to induce users to visit fake websites to get sensitive information like their private account, token for payment, credit card information, and so on.

The first phishing attack on AOL (America Online) can be traced back to early 1995 [6]. A phisher successfully obtained AOL users personal information. It may lead to not only the abuse of credit card information, but also an attack on the online payment system entirely feasible.

The phishing activity in early 2016 was the highest ever recorded since it began monitoring in 2004. The total number of phishing attacks in 2016 was 1,220,523. This was a 65 percent increase over 2015. In the fourth quarter of 2004, there were 1,609 phishing attacks per month. In the fourth quarter of 2016, there was an average of 92,564 phishing attacks per month, an increase of 5,753% over 12 years [7]. According to the 3rd Microsoft Computing Safer Index Report released in February 2014, the annual worldwide impact of phishing could be as high as $5 billion [8]. With the prevalence of network, phishing has become one of the most serious security threats in modern society, thus making detecting and defending against web phishing an urgent and essential research task. Web phishing detection is crucial for both private users and enterprises [9].

Some possible solutions to combat phishing were created, including specific legislation and technologies. From a technical point of view, the detection of phishing generally includes the following categories: detection based on a black list [10] and white list, detection based on Uniform Resource Locator (URL) features [11], detection based on web content, and detection based on machine learning. The antiphishing way using blacklist may be an easy way, but it cannot find new phishing websites. The detection on URL is to analyze the features of URL. The URL of phishing websites may be very similar to real websites to the human eye, but they are different in IP. The content-based detection usually refers to the detection of phishing sites through the pages of elements, such as form information, field names, and resource reference.

In this paper, we will focus on the detection model using a deep learning framework. The main contributions are as follows:(i)We present two feature types for web phishing detection: an original feature and an interaction feature. The original feature is the direct feature of URL, including special characters in URL and age of the domain. The interacting feature is the interaction between websites, including in-degree and out-degree of URL.(ii)We introduce DBN to detect web phishing. We discuss the training process of DBN and get the appropriate parameters to detect web phishing.(iii)We use real IP flows data from ISP to evaluate the effectiveness of the detection model on DBN. True Positive Rate (TPR) with different parameters is analyzed; our TPR is approximately 90%.

The paper is organized as follows.

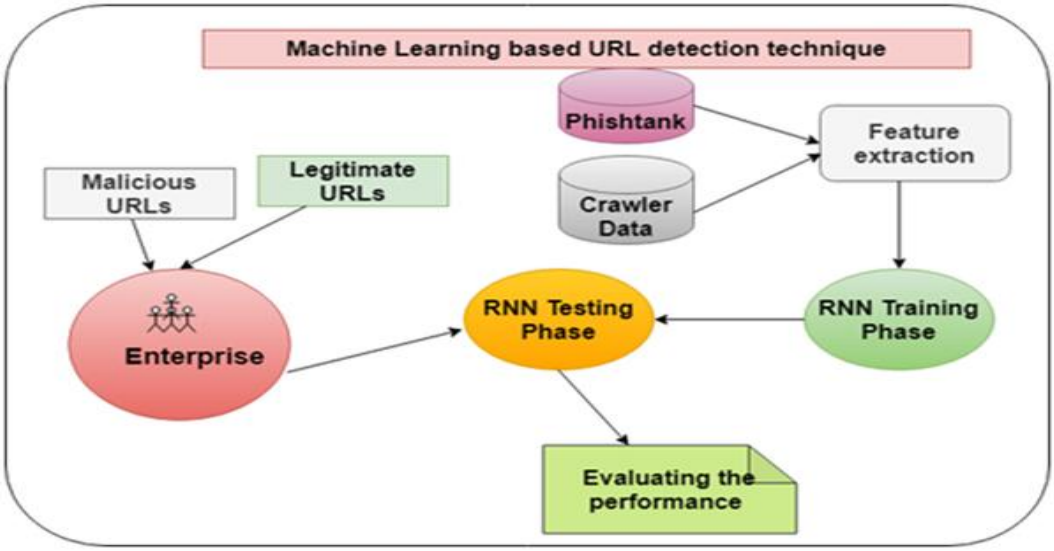
1.Related works are discussed in Section

2. The detection model and algorithm are discussed in Section

3. DBN is tested and evaluated in Section

4. The conclusion is drawn in Section

DIAGRAM



Methodology:

The methods of detecting phishing attacks are:

Google Safe Browsing:

This approach uses the blacklist URLs to discover the phishing attack. A sample URL is taken as input and checked within the blacklist repository. If the URL is present in the black list repository, the URL is termed as suspicious URL, else it is a legitimate website. The main shortcoming of this approach is its inability to detect the phishing URL which aren't present within the blacklist which could increase the false positives rate.

Spoof Guard:

This method scans suspicious websites for phishing symptoms to determine whether the website is legitimate or phishing. Some heuristics include image verification, link verification, URL verification and password field verification. If the total score of the phishing symptoms listed above exceeds the threshold, it is classified as a legitimate phishing website. This method detects zero-day attacks. This method also has a high limit on the number of false positives.

False alert: This method will use visual phishing detection when the attacker uses the same CSS style to deceive the original website. In this method, CSS style comparisons are performed on white listed websites with suspicious website styles to detect phishing. attack.

Method	Disadvantages
Early detection and manual blocking of phishing sites.	Most Internet users do not know how to identify phishing websites in real time. Even experienced people are often attacked for forgetting to check the legitimacy of the website. They do not provide safety training when they are busy at work.
Detection of website content and URL [4]	T he URL detection of new website is insufficient. These methods are not precise and usually produce a small number of false positives
Block the phishing E-Mails by various spam filter software [3]	These spam filters tend to block genuine messages. They fail to find these attacks excluding from email-threads
Server – side Detection	Users can receive delayed responses from servers concerning the credibility of the web site. They underperform in slow internet connections. Client – side Detection These software's signature - based security control
Client – side Detection	These software's signature - based security controls are proving less and less effective as years pass by. For example, these solutions are not particularly good at identifying file – less malware. They utilize a lot of memory.
Other Detection Methods	It is not effective on pages that are not visited previously and websites should be maintained by constantly updating to preserve better accuracy.

References

- Dilbag Singh, Saloni Manhas, and Swapnesh Taterh “A Novel based Approach for Phishing Websites Detection using Decision Tree” [1]

  By making use of decision tree algorithm to classify information gain, financial gain and other

  uncertainty FST to increase the performance of anti-phishing detection application. In order to overcome issues of phishing attack, anti-phishing detection was designed to detect phishing website URLs on the victim’s email. Besides that, anti-phishing detection application able to generate a report of phishing website which are attached on victim’s email.

- Jitendra Kumar, A. Santhanavijayan, B. Janet, B.S. Bindhumadhava, and Balaji Rajendran “Phishing Website Classification and Detection Using Machine Learning” [2]

  By making use of lexical structure od URL to classify url into different parts and identify the Url whether the given url is phishing url or not. In, this paper, they have compared different machine learning techniques for the phishing URL classification task and achieved the highest accuracy of 96% for Naïve Bayes Classifier with a precision=1, recall = .95 and F1-Score= .96.

- Chua Shang Ren; Rabab Alayham Abbas Helmi; Muhammad Irsyad Abdullah; Arshad Jamal “Email Anti-Phishing Detection Application” [3]

  There are many techniques to overcome tricked by phishing website. One of the methods mostly used to detect phishing is by using visual similarity. This method is to dissimilar phishing webpage, which also reduce the successful rate of victim got tricked by phishing scams. Besides that, there is another method to detect phishing website is by using compression algorithm. Compression algorithm is a critical component which perform a compression of nine compressors that include 1-dimensional string and 2- dimensional image compression. The main aim of this paper is to spot phishing attacks that connects the victim’s email by mistreatment by applying decision tree algorithm that enforced within the application. This project mainly focused on detect on attachment file of phishing website in the email buddies by using

decision tree algorithm. Anti-Phishing detection application used to detect, identify and block the phishing website or email that effected by the phishing website. It is able to calculate the percentages of stored phishing emails in the user's email.

- Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh, Aram Alsedrani "Detecting Phishing Websites Using Machine Learning" [4]

  The system acts as an extra functionality to a web browser as extension that mechanically notifies the user once it detects a phishing web site. The system is predicated on a machine learning method, notably supervised learning. They've got selected the Random Forest technique because of its sensible performance in classification. The focus will be on the features combination that we get from Random Forest (RF) technique, as it has good accuracy, is relatively robust, and has a good performance. Recently, there have been several studies that are trying to solve the phishing problem. They can be classified into four types: blacklist, heuristic, content analysis, and machine learning techniques. The blacklisting technique compares the URL with an existing database that contains a list of phishing website URLs. Because of the rapid increase of such phishing attacks, the blacklist approach has become more inefficient in checking whether each URL is a phishing website or not, and this kind of delay can also lead to zero-day attacks from these new phishing sites.

- Fuma Dobashi, Akihito Nakamura, "Proactive Phishing Sites Detection,"[5]

  In this paper, emphasized compared phishing mitigation techniques, such as blacklist, heuristics, visual similarity, and machine learning and concluded that these techniques have limitations in dealing with zero-hour attacks and proactive detection of phishing websites. The authors proposed suspicious URL's generation and to predicts likely phishing sites from the given legitimate brand domain name and scores and judges suspects by calculating various indexes to detect phishing websites.

- Akbar-Siami Namin, Moitrayee Chatterjee, "Detecting Phishing Websites through Deep Reinforcement Learning"[6]

  This paper proposed a deep reinforcement learning model to detect malicious URLs. This model is capable of adapting to the dynamic behavior of the

phishing sites and thus even learn the features associated with phishing website detection. The proposed model uses Deep Reinforcement Learning Techniques. They got an accuracy of 90%.

- Dimitris Tsaptsinos; Martyn Weedon; James Denholm-Price “Random Forest explorations for URL classification”[7]

  This paper builds the classifier using Random Forest techniques. These RF techniques are used to classify the given url into substring and then consider them. In this paper, the main objective is to evaluate the performance of the Random Forest algorithm using a lexical only dataset. The performance is benchmarked against some other machine learning techniques and additionally against those reported in the literature. Initial results from experiments indicate that the Random Forest algorithm performs the best yielding an 86.9% accuracy.

LITERATURE REVIEW

In this Review, Many Papers have studied to know the details about web phishing detection. Machine learning, classification algorithm and other technique can be involved. here, explain identification techniques of each paper.

SURVEY PAPERS

Rao et al proposed a novel classification approach that use heuristic based feature extraction approach. In this, they have classified extracted features into three categories such as URL Obfuscation features, Third-Party-based features, Hyperlink-based features. Moreover, proposed technique gives 99.55% accuracy. Drawback of this is that as this model uses third-party features, classification of website dependent on speed of third-party services. Also this model is purely depends on the quality and quantity of the training set and Broken links feature extraction has limitation of more execution time for the websites with more number of links.

Chunlin et al. proposed approach that primarily focus on character frequency features. In this they have combined statistical analysis of URL with machine learning technique to get result that is more accurate for classification of

malicious URLs. Also they have compared six machine-learning algorithms to verify the effectiveness of proposed algorithm which gives 99.7% precision with false positive rate less than 0.4%.

Sudhanshu et al. used association data mining approach. They have proposed rule based classification technique for phishing website detection. They have concluded that association classification algorithm is better than any other algorithms because of their simple rule transformation. They achieved 92.67% accuracy by extracting 16 features but this is not up to mark so proposed algorithm can be enhanced for efficient detection rate.

M. Amaad et al. presented a hybrid model for classification of phishing website. In this paper, proposed model carried out in two phase. In phase 1,they individually perform classification techniques, and select the best three models based on high accuracy and other performance criteria. While in phase 2, they further combined each individual model with best three model and makes hybrid model that gives better accuracy than individual model. They achieved 97.75% accuracy on testing dataset. There is limitation of this model that it requires more time to build hybrid model.

Hossein et al. developed an open-source framework known as "Fresh-Phish". For phishing websites, machine-learning data can be created using this framework. In this, they have used reduced features set and using python for building query .They build a large labelled dataset and analyse several machine-learning classifiers against this dataset .Analysis of this gives very good accuracy using machine-learning classifiers. These analyses how long time it takes to train the model.

Gupta et al. proposed a novel anti phishing approach that extracts features from client-side only. Proposed approach is fast and reliable as it is not dependent on third party but it extracts features only from URL and source code. In this paper, they have achieved 99.09% of overall detection accuracy for phishing website. This paper have concluded that this approach has limitation as it can

detect webpage written in HTML .Non-HTML webpage cannot detect by this approach.

Bhagyashree et al. proposed a feature based approach to classify URLs as phishing and non-phishing. Various features this approach uses are lexical features, WHOIS features, Page Rank and Alexa rank and Phish Tank-based features for disguising phishing and non-phishing website. In this paper, web-mining classification is used.

Mustafa et al. developed safer framework for detecting phishing website. They have extracted URL features of website and using subset based selection technique to obtain better accuracy .In this paper, author evaluated CFS subset based and content based subset selection methods And Machine learning algorithms are used for classification purpose.

Priyanka et al. proposed novel approach by combining two or more algorithms. In this paper ,author has implemented two algorithm Adaline and Backpropion along with SVM for getting good detection rate and classification purpose. Pradeepthi et al.[15] In this paper ,Author studied different classification algorithm and concluded that tree-based classifier are best and gives better accuracy for phishing URL detection. Also Author uses variousfeatures such as lexical features, URL based feature, network based features and domain based feature.

Luong et al. proposed new technique to detect phishing website. In proposed method, Author used six heuristics that are primary domain, sub domain, path domain, page rank, and alexa rank, alexa reputation whose weight and values are evaluated. This approach gives 97 % accuracy but still improvement can be done by enhancing more heuristics.

Ahmad et al. proposed three new features to improve accuracy rate for phishing website detection. In this paper, Author used both type of features as commonly known and new features for classification of phishing and non-phishing site. At the end author has concluded this work can be enhanced by using this novel features with decision tree machine learning classifiers.

Mohammad et al. proposed model that automatically extracts important features for phishing website detection without requiring any human intervention. Author has concluded in this paper that the process of extracting feature by their tool is much faster and reliable than any manual extraction.

Oluwatobi Ayodeji Akanbi, ... Elahe Fazeldehkordi, in A Machine-Learning Approach to Phishing Detection and Defense, 2015

SUMMARY REVIEW:

Thus to summarize, we have seen how phishing is a huge threat to the security and safety of the web and how phishing detection is an important problem domain. We have reviewed some of the traditional approaches to phishing detection; namely blacklist and heuristic evaluation methods, and their drawbacks. We have tested two machine learning algorithms on the 'Phishing Websites Dataset' and reviewed their results. We then selected the best algorithm based on it's performance and built a Chrome extension for detecting phishing web pages. The extension allows easy deployment of our phishing detection model to end users. We have detected phishing websites using Random Forest algorithm with and accuracy of 97.31%. For future enhancements, we intend to build the phishing detection system as a scalable web service which will incorporate online learning so that new phishing attack patterns can easily be learned and improve the accuracy of our models with better feature extraction.

We analyze the features of phishing websites and present two types of feature for web phishing detection original feature and interaction feature. Ten we introduce DBN to detect phishing websites and discuss the detection model and algorithm for DBN. We train DBN and get the appropriate parameters for detection in the small data set. In the end, we use the big data set to test DBN and TPR is approximately 90%.

The importance to safeguard online users from becoming victims of online fraud, divulging confidential information to an attacker among other effective uses of phishing as an attacker's tool, phishing detection tools play a vital role in ensuring a secure online experience for users. Unfortunately, many of the existing

phishing-detection tools, especially those that depend on an existing blacklist, suffer limitations such as low detection accuracy and high false alarm that is often caused by either a delay in blacklist update as a result of human verification process involved in classification or perhaps, it can be attributed to human error in classification which may lead to improper classification of the classes. These critical issues have drawn many researchers to work on various approaches to improve detection accuracy of phishing attacks and to minimize false alarm rate. The inconsistent nature of attacks behaviors and continuously changing URL phish patterns require timely updating of the reference model. Therefore, it requires an effective technique to regulate retraining as to enable machine learning algorithm to actively adapt to the changes in phish patterns.

This study focus on investigating a better detection approach and to design an ensemble of classifier suitable to be used in phishing detection. Figure 6.1 summarizes the design and implementation phases leading to the proposed better detection model.

Fig. 6.1. Design and development phases leading to the proposed model.

Phase 1 focuses on data set gathering, pre processing, and feature extraction. The objective is to process data for use in Phase 2. The gathering stage is done manually by using Google crawler and Phishtank, each of this data gathering methods were tested to ensure a valid output. The data set is validated first after gathering, then normalized, features extraction and finally dataset division. Nine features were selected for this project to ensure an optimum result from the classifiers and also, since using a small feature set will invariably speed up processing time for training and for classification of new instances. These features were selected on the basis of the weighted performance of each feature by using information gain algorithm to ensure that only the best features were selected. This phase focuses on ensuring that the dataset pre processing is done appropriately to accommodate the models selected.

Phase 2 focuses on design and implementation of training and validating model using single classifier. A predefined performance metrics is used as a measurement of accuracy, precision, recall, and f-measure. The objective of this phase is to test the performance of individual classifiers in the pool of varying dataset as divided in Chapter 4 and select the most performed of all the reference classifiers. An accuracy of 99.37% was obtained from K-NN which is the highest as compared to other classifiers referenced. Although it was also observed that some of the classifiers like K-NN and C4.5 maintained a close range performance, same cannot be said of the remaining two classifiers that appeared lacking behind in performance. The performance of K-NN is not surprising since the dataset used is of a small set and as such K-NN often perform better with small dataset but the performance decreases has the size of the dataset increases (Kim and Huh, 2011). Also, since the performance of KNN is primarily determined by the choice of K, the best K was found by varying it from 1 to 7; and found that KNN performs best when K = 1. This as well, helped in the high accuracy of KNN compared to other classifiers used.

Phase 3 which corresponds to the third objective is divided into two parts, one is the ensemble design and the other is the comparative study between the best ensemble and the best individual classifier that was selected in Phase 2. To design a good ensemble, only three algorithms are used for individual ensemble due to the selection of majority voting as the ensemble algorithm, odd number of algorithms must be used to select the committee of ensembles. For every instance of each ensemble, an ensemble design of three algorithms is being selected until all the algorithms have been combined evenly. The design ensemble performed very well with an accuracy of 99.31% for the best-performed ensemble and this result is then compared with that obtained in Phase 2. The outcome of the comparison suggests that if K-NN algorithm is removed or if the size of the data set is increased, the ensemble will most likely perform better than the individual algorithm. This investigation will be considered as part of future work.