A Synopsis on

Heuristic Based Approach for Phishing Site Detection

Submitted in partial fulfillment of the requirements
of the degree of

Bachelor of Engineering

in

Information Technology

by

Chirag L. Chaudhari (15204030)
Swapnil S. Ghawali (15204009)
Swapnil R. Kshetre (15204003)
Aakash A. Sane (15204012)

Under Guidance of Prof. Sunil A. Sushir

Department of Information Technology
A.P. Shah Institute of Technology
G.B.Road,Kasarvadavli, Thane(W), Thane-400 615
UNIVERSITY OF MUMBAI
2018-2019

CERTIFICATE

This is to certify that the project Synopsis entitled ***"Heuristic Based Approach for Phishing Site Detection"*** Submitted by ***"Chirag Lahu Chaudhari (15204030), Swapnil Shivaji Ghawali (15204009), Swapnil Ravsaheb Kshetre(15204003), Aakash Anil Sane(15204012)"*** for the partial fulfillment of the requirement for award of a degree ***Bachelor of Engineering*** in ***Information Technology*** to the University of Mumbai,is a bonafide work carried out during academic year 2018-2019

Prof. Sunil A. Sushir
(Guide)

Prof. Kiran Deshpande
(Head Department of Information Technology)

Dr. Uttam D.Kolekar
(Principal)

External Examiner(s)

1.

2.

Place:A.P.Shah Institute of Technology, Thane
Date:

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, We have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

_____

(Signature)
(Chirag Lahu Chaudhari [15204030])

_____

(Signature)
(Swapnil Shivaji Ghawali [15204009])

_____

(Signature)
(Swapnil Ravsaheb Kshetre [15204003])

_____

(Signature)
(Aakash Anil Sane [15204012])

Date:

Abstract

Phishing has been a major security threat in which there is a huge loss for companies as well as customers. These phishing attacks are increasing day by day due to lack of efficient detection techniques and effective preventive measures. This paper proposes a Heuristic-based phishing detection algorithm. In particular, this research focuses on improving upon the previously published text-based approach. The algorithm in the previous work analyzes the body text in an email to detect whether the email message asks the user to do some action such as clicking on the link that directs the user to a fraudulent website .

Introduction

With the recent growth of the Internet environment and diversification of available web services, web attacks have increased in quantity and advanced in quality. Phishing is a type of social engineering attack that targets a users sensitive information through a phony website that appears similar to a legitimate site, or by sending a phishing email[1].

Phishing is a website forgery technique with an intention to track and steal the sensitive information of online users. The hacker fools the user with social engineering techniques such as SMS, voice,email, website and malware[2].Various approaches have been proposed and implemented to detect a variety of phishing attacks such as use of blacklists and whitelists to name a few. We propose a desktop application called PhishSaver, which focuses on URL and website content of the phishing web page. We aim at detecting phishing websites with the help of a desktop application named PhishSaver. Phish-Saver use a heuristic features to detect a number of phishing attacks[2].

We will create a Application which name is PHISHSERVER. PhishSaver is a desktop application to effectively detect phishing websites and practices. PhishSaver is capable to detect most of the phishing techniques and aims at providing full-proof security from all kinds of phishing attacks.PhishSaver is not a traditional anti-virus software and does not guarantee any protection from any kind of virus attacks. PhishSaver is based on the idea that users should be able to browse the internet safely and access websites without getting concerned about the legitimacy of the websites. The system works by taking an input from the user in the form of URL of the website for which legitimacy needs to be determined. The system then outputs the state of the website as phishing, legitimate or unknown[5].

The existing systems have several limitations which can be overcome by PhishSaver. The main advantages of PhishSaver over normal phishing detectors are as following:

1. It is based on heuristic approach that is capable of detecting Zero hour phishing attacks that is phishing attacks that are relatively new which is not possible for most of the other phishing detectors.

2. Users just need to provide the URL of the website whose legitimacy needs to be determined. Nothing else needs to be done by the user.

3. The main advantage of our application is that it can detect phishing sites which tricks the users by replacing content with images, which most of the existing anti phishing techniques are not able to detect, even if they can, they take more execution time than our application.

Heuristic based methods extract features of a web page to decide the legitimacy of the website in-stead of depending on any precompiled lists. Most of these features are extracted from URL and HTML Document Object Model (DOM) of the given web-page. The extracted features are com-pared with known features collected from phishing and legitimate pages to decide its legitimacy. Some of these approaches use heuristics to calculate spoof score of a given web page to check its genuineness[5]

Machine learning approach :- The machine learning approach exploits many characteristics of the URL and the websites by using machine learning techniques such as Support Vector Machines (SVM), Decision tree algorithms, Random forest classification method etc. These characteristics are combined to use to detect the phishing websites. However, there are some limitations in this approach.First, the machine learning-based techniques might fail in the case that attackers compromise legitimate domains and host phishing attacks on those servers. Second, because of text-based analysis mechanism, these phishing detection techniques cannot detect the phishing websites which are purely made up of images[7].

Heuristics based Phishing Detection :- Heuristic based methods extract features of a web page to decide the legitimacy of the website in-stead of depending on any precompiled lists. Most of these features are extracted from URL and HTML Document Object Model (DOM) of the given web page. The extracted features are com-pared with known features collected from phishing and legitimate pages to decide its legitimacy. Some of these approaches use heuristics to calculate spoof score of a given web page to check its genuineness[7]

Decision tree - Machine learning technique:- Machine learning algorithms are used to build an efficient classifier which would decide whether a given URL is phishing or not. Decision tree is a classification method that was introduced in 1992 by Quinlan[10] . It creates a tree form for classifying samples. Each internal node of the tree corresponds to a feature, and the edges from the node separate the data based on the value of the feature [10]. Decision tree includes a decision area and leaf node. The decision area checks the condition of the samples and separates them into each leaf node or the next decision area. The decision tree is very fast and easy to implement; however, it has the risk of overfitting we propose a new heuristic-based phishing detection technique that resolves the limitation of the blacklist-based technique. Even complex phishing attacks can be easily determined by Decision Tree algorithm. This algorithm has relatively less false positive and false negative rates[10].

We implemented the proposed technique and conducted an experimental performance evaluation. The proposed technique extracts features in URLs of user-requested pages and applies those features to determine whether a requested site is a phishing site. This technique can detect phishing sites that cannot be detected by blacklist-based techniques; therefore, it can help reduce damage caused by phishing attacks.

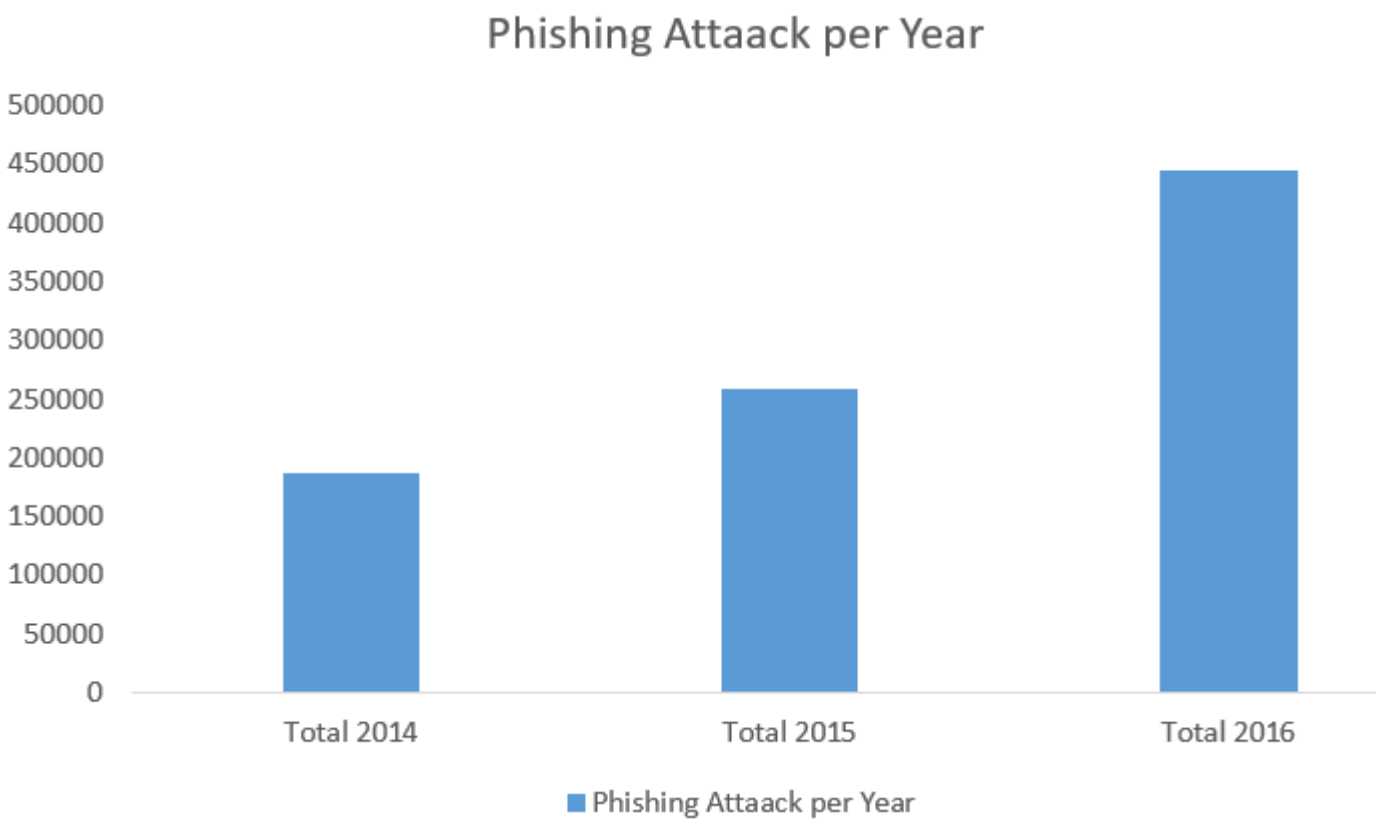

Figure 1: Phishing Attack Per Year

Objectives

In this project URL heuristic approach will be used along with the ranking of sites to extract the features from the URL. All the extracted features along with the phishing and legitimate sites URLs will be stored in database. Next, a classifier will be generated using decision tree algorithm which will classify the URLs as phishing and legitimate.When new URL is received it will extract the features and will compare them with the features stored in database, thus classifying the incoming site as phishing or legitimate.

Literature Review

Title	Problem Identified	Methodology	Strength	weakness
The Sceured Anti-Phishing Approach Using image based Validation. Y. Yesu Jyothi, D.Srinivas & k.GovindaRaju,2013	To solve the problem of phishing & protect individual personal private information	Visual Cryptography (image based validation)	It prevents attack of phishing websites on financial portal,banking portal & online shopping market	Inability to recover missing or corrupt share
Protecting users Against phishing attacks. Engin kirda & Chistopher Kruegel,2012	Increased email linked to phishing scams	Browser Extension	It protect users against spoofed website – based phishing attacks	It requires that user support to capture & store sensitive information rather than automatically captureing & storing the sensitive information
Phishnet Anti-Phishing Technique(Prakash et al.,2010)	Predicts varation of URLs	Heuristics	It replace Top level domain(TLD),Dirctory structure similarity,IP address equivalence,Qury string substitution & brand name equivalence	It connot detect zero day phishing
RDF based anti phishing framework(Vamsee et al.,2013)	To differentiate a phishing site from a legitimate site	Resource Description framework model	It uses nineteen properties that describes the characteristics of a webpage to distinguish between a legitimate & a phishin site	Building RDF can be cumbersome

Problem Definition

A comprehensive efficient detection technique should be developed in order to detect and inform the web users about the phishing attacks to make sure that their sensitive data will not be disclosed during these attacks. This research project deals with a comprehensive heuristic based method for phishing detection which is based on content of the website through which phishing attacks can be discovered.

Proposed System Architecture/Working

The proposed architecture for System is given in Figure 2. The system provides an interface where the user can write his/her query. Once the search button is clicked, it gives the list of URLs on the same page . A URL is a protocol that is used to indicate the location of data on a network. The URL is composed of the protocol, sub domain, primary domain, top-level domain (TLD), and path domain. In meantime; it saves all the URLs in the database. The protocol refers to a communication protocol for exchanging information between information devices; e.g., HTTP, FTP, HTTPS, etc. Protocols are of various types and are used in accordance with the desired communication method. For each URL, all mentioned eighteen factors are

calculated and saved as total score in the database. Then based on the total score value of each URL, they are rearranged in the descending order, which means if URL has high total score value then it will appear as the top most result and accordingly rest comes as per their total score value.

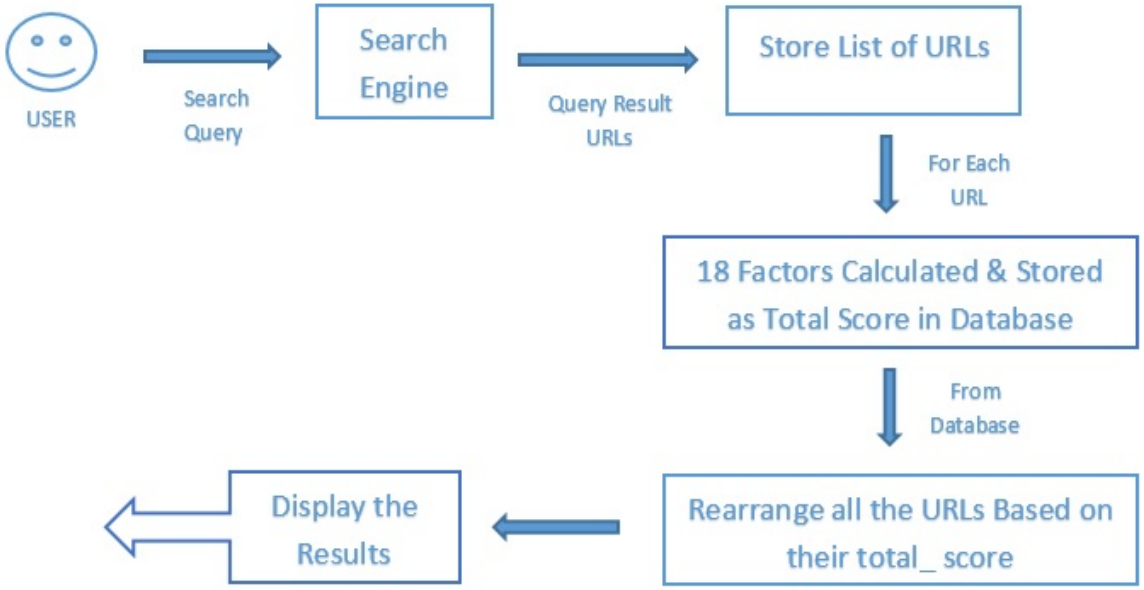

Figure 2: System Diagram

Summary

The proposed model can reduce damage caused by phishing attacks because it can detect new and temporary phishing sites. System also implemented decision tree algorithm and generated tree for it We will be looking forward for the new features to use and try to improve more accuracy and reduce false positive value of the system. We also look forward to discover new feature with high impact to detect phishing. Also make plugin for the browser which will alert user about phishing website and reduce damage cause with it as much as possible.

References

[1] Khonji, Mahmoud, Youssef Iraqi, and Andrew Jones."Phishing detection: a literature survey."Communications Surveys Tutorials,IEEE 15.4(2013): 2091-2121.

[2] APWG, Phishing activity trends paper.[online]. http://docs.apwg.org/reports/APWG Global Phishing Report 1H 2014.pdf

[3] Luong Anh Tuan Nguyen1, Ba Lam To1, Huu Khuong Nguyen1 and Minh Hoang Ngu-yen21 Faculty of Information Technology, 2014 IEEE An Ecient Approach for Phishing Detec-tion Using Single-Layer Neural Network

[4] So Young Rieh,Judgment of Information Quality and Cognitive Authority in the Web,citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.107.8991

[5] Suman Bhattacharyya, Chetan kumar Pal, Praveen kumar Pandey, Detecting Phishing Websites, a Heuristic Approach,International Journal of Latest Engineering Research and Applications (IJLERA) ISSN: 2455-7137 Volume 02, Issue 03, March 2017, PP 120-129

[6] Nguyen, Luong Anh Tuan, et al. "A novel approach for phishing detection using URL-based heuristic." Computing, Management and Telecommunications (ComManTel), 2014International Conference on. IEEE, 2014

[7] Wikipedia. (2015. March) Uniform Resource Loactor.Avaliable: http://en.wikipedia.org/wiki/Uniform resource locator

[8] Sunil, A. Naga Venkata, and Anjali Sardana. "A pagerank based detection technique for phishing web sites." Computers Informatics (ISCI), 2012 IEEE Symposium on. IEEE, 2012.

[9] WANG, Wei-Hong, et al. "A Static Malicious Javascript Detection Using SVM." strings. Vol. 40. 2013

[10] Hou, Yung-Tsung, et al. "Malicious web content detection by machine learning." Expert Systems with Applications 37.1 (2010): 55-60.