DETECTING PHISHING WEBSITE USING ASSOCIATIVE CLASSIFICATION

Shubham Walde[1], Piyush Katkar[2], Sneha Waghmare[3], Snehal Dhapodkar[4], Aishwarya Badole[5],
Prof. P. S. Moon
Department of Computer Technology
K.D.K. College of Engineering, Nagpur, India

ABSTRACT: Phishing scam is known unlawful activity in which victims are defraud to disclose the confidential data and information specially related to user personal financial information. There are various phishing schemes such as deceptive, malware based, DNS-based, etc. Hence in this paper, a systematic review analysis on existing works related with the phishing detection and response techniques together with apoptosis have been further investigated and evaluated. Phishing is a significant problem involving fraud email and web sites that mislead unsuspecting users into disclosing private information. In this paper, we present the design, implementation, and evaluation of various techniques for detecting phishing web sites. Phishing websites are fake websites that are created by dishonest people to mimic web pages of real websites. Victims of phishing attacks may expose their financial sensitive information to the attacker whom might use this information for financial and criminal activities. This paper investigates features selection aiming to determine the effective set of features in terms of classification performance.

Keywords: CANTINA, Data Mining, phishing websites, Security, website security.

I. INTRODUCTION

As online financial activities are on the rise, so have online illegal activities in which phishing is playing a major role for illegally obtaining private individual details. Thus the online technology is growing as faster level, so have other numerous online activities such as advertising, gaming, and e-commerce etc are growing at faster rate. Phishing activities against financial institutions have become a regular occurrence leading to a rising concern about how to increase security on these sectors which could relate to banks and online shopping. There are some examples such as EBay and Amazon. This illegal schemes are conducted via the Internet are generally difficult to trace and carry on, hence they cost individuals and businesses millions of dollars every year. From computer viruses to web site hacking and financial fraud to Internet crime became a larger concern than ever in the late 1990s and early 2000s. In response to such issues, there are some different anti-phishing tools were developed in order to counter such illegal online activities.

Phishing emails are also contain some links which contains the affected website where they asked to type the personal information of user such as username and password of account details, so that the website will hack the typed data by user. As phishing is increasing day by day there are many tricks developed against the existing anti-phishing. Phishing email are also sent to a many people and the phishers will also try to count the percentage of people who read that email and entered the information. As a result researchers are attempting to reduce the risk and vulnerabilities of such illegal phishing activities. Some researchers also define phishing as a new type of network attack. The attacker develops duplicates of existing Web sites to fool users. It is very difficult to find that the individuals are actually visiting an actual site or malicious site. Phishing is also understood to be a sort of brand spoofing or carding. for example by using specially designed e-mails or instant messages into submitting private, financial, or password data to what they think is their service provides' Website.

i. OBJECTIVE

Latest Technique of Detecting Phishing Website, this will help users to verify that their browser has made a secure connection to a trusted site. Possible to detect phisher, even if a user falls for a phishing site, the phishers would not see the correct password Users. There are always innovative ways that are created regularly by phishing attackers to confuse the anti-phishing techniques the main aim is to enhance the security of websites.Phishing itself is not a new concept, but it's increasingly used by phishers to steal user information and perform business crime in recent years,preventing users from submitting personal information is the main aim. Making user payment safely for better transactions is also an objective of this phishing problem.

ii. LITERATURE SURVEY

- Phishing Detection using Content Based Associative Classification Data Mining [1]. In this paper it is intended to prevent a phishing using data mining technique. MCAC Algorithm is gives higher efficiency towards to detect phishing activity. In MCAC algorithm does not consider a content based features of websites. It is intended to add content and page style features in that algorithm and change the system for better performance. This paper shows proposed method and flow chart. This paper also shows all the features of website which are considered during experimental analysis.
- Content Based Approach for Detection of Phishing sites [2]. In this paper, we present the design, implementation, and evaluation of content-based approach to detecting phishing web sites. We also discuss the design and evaluation of several heuristics we developed to reduce false positives. Our experiments show that CANTINA is good at detecting phishing sites, correctly labelling approximately 95% of phishing sites.
- Phishing Websites Detection based on Phishing Characteristics in the Webpage Source Code [3]. In this paper, we propose a phishing detection approach based on checking the webpage source code, we extract some phishing characteristics out of the W3C standards to evaluate the security of the websites, and check each character in the webpage source code, if we find a phishing character, we will decrease from the initial secure weight. Finally we calculate the security percentage based on the final weight, the high percentage indicates secure website and others indicates the website is most likely to be a phishing website. We check two webpage source codes for legitimate and phishing websites and compare the security percentages between them, we find the phishing website is less security

percentage than the legitimate website; our approach can detect the phishing website based on checking phishing characteristics in the webpage source code.

- An Associative Classification Data Mining Approach for Detecting Phishing Websites [4]. This paper, proposes a new AC algorithm called Phishing Associative Classification (PAC), for detecting phishing websites. PAC employed a novel methodology in construction the classifier which results in generating moderate size classifiers. The algorithm improved the effectiveness and efficiency of a known algorithm called MCAR, by introducing a new prediction procedure and adopting a different rule pruning procedure.
- Detection and Prediction of Phishing Websites using Classification Mining Techniques [5]. This paper investigates features selection aiming to determine the effective set of features in terms of classification performance. We compare two known features selection method in order to determine the least set of features of phishing detection using data mining. Experimental tests on large number of features data set have been done using Information Gain and Correlation Features set methods. Further, two data mining algorithms namely PART and IREP have been trained on different sets of selected features to show the pros and cons of the feature selection process.
- Associative Classification Mining for Website Phishing Classification [6]. In this article, an Associative classification (AC) data mining algorithm that uses association rule methods to build classification systems (classifiers) is developed and applied on the important problem of phishing classification. The proposed algorithm employs a classifier building method that discovers vital rules that possibly can be utilised to detect phishing activity based on a number of significant website’s features. Experimental results using the proposed algorithms and three other rule based algorithms on real legitimate and fake websites collected from different sources have been conducted. The results reveal that our algorithm is highly competitive in classifying websites if contrasted with the other rule based classification algorithms with respective to accuracy rate.

II. PHISHING WEBSITE DETECTION

Financial and governmental institutes offer a variety of financial services to their clients. Online banking and online shopping become popular since the late 80’s. Nowadays, almost all banks around the globe offer many online services to their clients while online shopping became a major sector of the world economy. Phishing is a method to imitating an official websites or genuine websites of any organization such as banks, institutes social networking websites, etc. The word ‘Phishing ’Initially emerged in 1990s. The early hackers often use ‘ph’ to replace ‘f’ to produce new words in the hacker’s community, since they usually hack by phones. Phishing is a new word produced from ‘fishing’, it refers to the act that the attacker allure users to visit a faked Web site by Sending them faked e-mails (or instant messages), and stealthily get victim’s personal information such as user name, password, and national security ID, etc. Mainly phishing is attempted to theft private credentials of users such as username, passwords, PIN number or any credit card Details etc. Phishing is attempted by trained hackers or attackers. Another trend of approaches for detecting phishing websites relies on using a machine learning or data mining algorithm that recognize the phishing website based on a set of characteristics or features that are extracted from the website. The features are recognized by experts to be distinguishing characteristics of a phishing website (e.g.,uniform resource locater (URL), age of domain). According to these approaches, phishing is a pattern recognition problem that can be solved by chosen the “right” set of features and a “suitable” pattern discovery or recognition algorithm.

III. METHODOLOGY

We proposed this methodology for detecting phishing websites. The CANTINA is a content-based approach for detecting phishing websites, based on the term frequency and inverse document frequency (TF-IDF) information retrieval algorithm. CANTINA examines the content of the webpage to determine whether the site is phishing website or not.

CANTINA INCLUDED SOME RULES IN THIS PROPOSED METHODOLOGY.

Age of Domain

This tool is check whether the age of domain is greater than 12 months or not. Initially the phishing sites lifespan is now 4.5 days but this tools does not account for phishing sites based on existing web sites where criminals have broken into the web server, nor does it account for phishing sites hosted on otherwise legitimate domains.

Suspicious URL

This tool check that this pages URL contains any of this symbol ‘@’ or ‘-‘, because ‘@’ symbol in the URL indicates that the string in its left side can be relinquish and consider only right part of the string after the symbol. An ‘-‘symbol is rarely used in the legal websites.

Suspicious Links

This tool checks if the given links in the page satisfies the above condition or not. If it satisfies the condition then its marked as suspicious link.

IP Address

It check if the given URL contains any IP address in its domain or not.

Images

All images in the website including the website logo should load from the same URL of the website and not from another website, where all links should be internal links not external links. Therefore, we check the links to detect any external links inside the source code.

IV. BASIC CONCEPT

TF - IDF stands for Term Frequency - Inverse Document Frequency, and the TF-IDF weight is a weight often used in information retrieval and text mining.

Typically, the TF-IDF weight is composed by two terms:

TF: Term Frequency

The Term Frequency measures how a term that occurs in a document regularly. Since every document which is unique in length, it is possible that a term would visible more times in documents than small ones. Thus, the term frequency is calculated as given below:

TF (b) = (Number of times term b appears in a document) / (Total number of terms in the document).

IDF: Inverse Document Frequency

The IDF measures that how a term is important in the document. At that time while calculating TF, all terms that are examine equal important. However it’s well known that certain terms that “of", "is ", and "that", become available in the document but they don’t have any much more importance. The IDF that are calculate using following way:

IDF (b) = log_e (Total number of documents / Number of documents with term b in it).

V. WHOIS:

WHOIS (pronounced as the phrase whois) is a query and response protocol that is widely used for querying databases that store the registered users or assignees of an Internet resource, such as a domain name, an IP address block, or an autonomous system, but is also used for a wider range of other information. This protocol stores and delivers database content in a human-readable format.

Locating Phishing Server:

- URL is nothing but IP Address.
- Using IP address our system will locate phishing server.

VI. PLANNING

Architecture Diagram

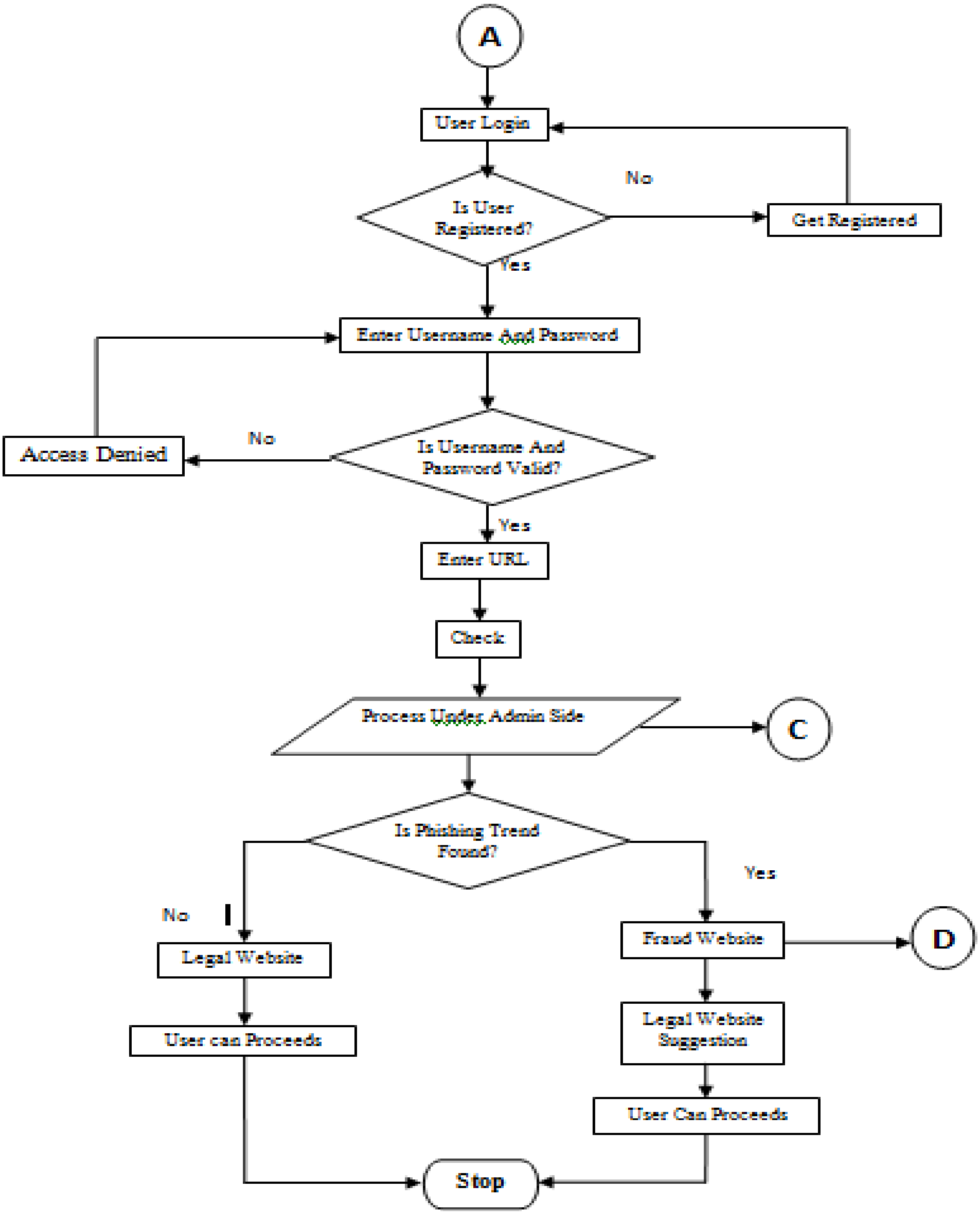

Fig. 6.1 – User Section

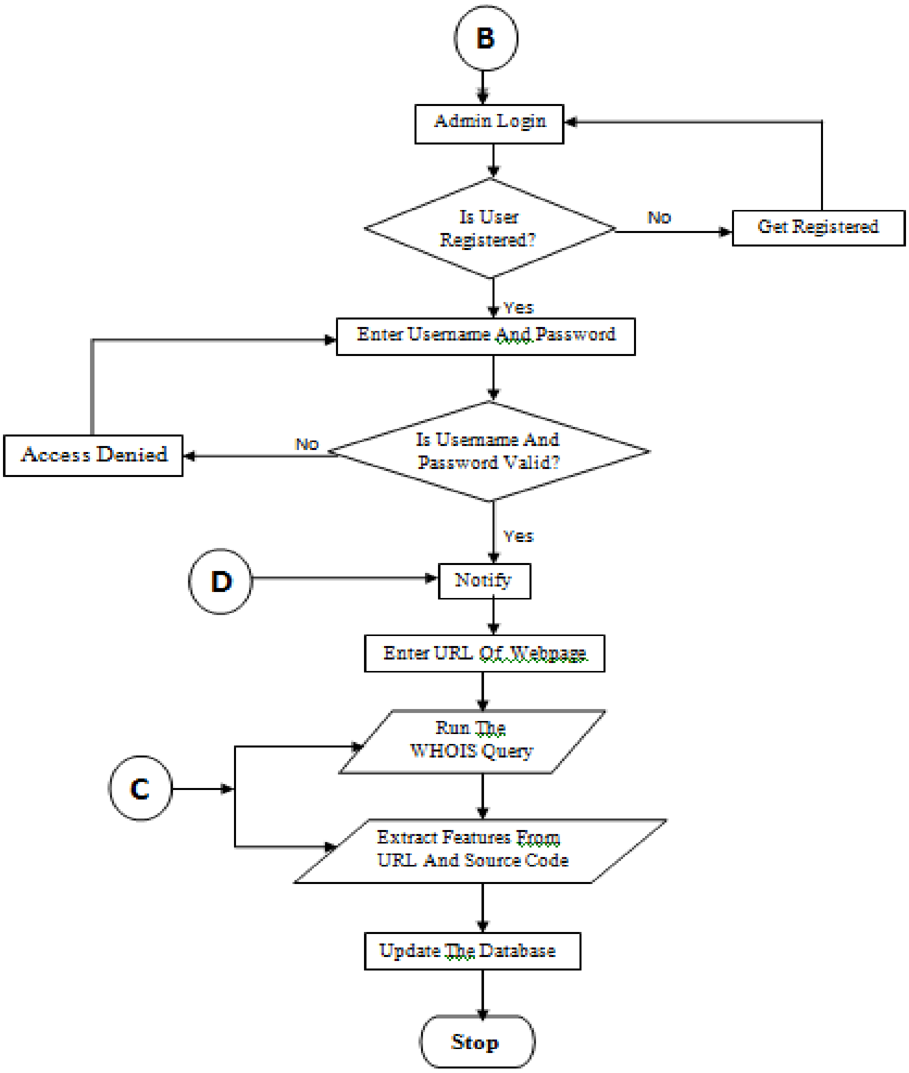

Fig. 6.2 – Admin Section

VII. CONCLUSION

Phishing is a significant problem involving fraudulent email and web sites that trick unsuspecting users into revealing private information. Here, we present the design, implement, and evaluated the CANTINA and TF-IDF techniques for detecting phishing web sites. If first module i.e. user module has been put in the work and required changes has implemented.

VIII. ACKNOWLEDGEMENT

The author would be like to thank Prof. P. S. Moon of K.D.K. College of Engineering, Nagpur, for her guidance, support and contribution for the success of this project and also for her invaluable assistance providing and hosting the necessary collaborative tools that made this project possible.

REFFERENCES

[1] Mitesh Dedakia, Khushali Mistry,” Phishing Detection using Content Based Associative Classification Data Mining”, Journal of Engineering Computers & Applied Sciences(JECAS) ISSN No: 2319-5606 Volume 4, No.7, July 2015.

[2] Anjali Gupta, Juili Joshi, Khyati Thakker, Chitra bhole,” CONTENT BASED APPROACH FOR DETECTION OF PHISHING SITES”, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 02 Issue: 01 | Apr-2015 www.irjet.net p-ISSN: 2395-0072.

[3] Mona Ghotaish Alkhozae, Omar Abdullah Batarfi,” Phishing Websites Detection based on Phishing Characteristics in the Webpage Source Code”, Volume 1 No. 6, October 2014 ISSN-2223-4985 International Journal of Information and Communication Technology Research.

[4] Suzan Wedyan, Fadi Wedyan “An Associative Classification Data Mining Approach for Detecting Phishing Websites”, Vol. 4, No. 12 December 2015 ISSN 2079-8407 Journal of Emerging Trends in Computing and Information Sciences.

[5] Mofleh Al-diabat, “Detection and Prediction of Phishing Websites using Classification Mining Techniques”, International Journal of Computer Applications (0975 – 8887) Volume 147 – No.5, August 2016.

[6] Neda Abdelhamid, Aladdin Ayesh, Fadi Thabtah, “Associative Classification Mining for Website Phishing Classification”, Informatics Dept, De Montfort University, Leicester, LE1 9BH.