International Journal of Advanced Research in Science Management and Technology
ISSN(Online)-2454-4159 Volume 6 Issue 1, January 2020

"Phishing Websites Detection Using Machine Learning"

Birari Harshal[1], Shelkar Akshay[2], Thakare Akshay[3],Patil Harshwardhan[4], Prof.S.P.Chavan[5]
UG Student, Dept. of Computer Engg., Gangamai College of Engineering Nagaon, Dhule,M.S., India[1,2,3,4]
Assistant Professor, Dept. of Computer Engg., Gangamai College of Engineering Nagaon, Dhule, M.S.,India[5]

Abstract—Phishing is a common attack on credulous people by making them to disclose their unique information using counterfeit websites Phishing website is one of the internet security problems that target the human vulnerabilities rather than software vulnerabilities. It can be described as the process of attracting online users to obtain their sensitive information such as usernames and passwords. In this paper, we offer an intelligent system for detecting phishing websites. The system acts as an additional functionality to an internet browser as an web page based extension that automatically notifies the user when it detects a phishing website. The system is based on a machine learning method, particularly supervised learning.

Keywords— Phishing, Phishing Websites, Detection, Machine Learning.

I. INTRODUCTION

Phishing is the most unsafe criminal exercises in cyber space. Since most of the users go online to access the services provided by government and financial institutions, there has been a significant increase in phishing attacks for the past few years. Phishers started to earn money and they are doing this as a successful business.Web service is a communication protocol and software between two electronic devices over the Internet. Web services extends the World Wide web infrastructure to provide the methods for an electronic device to connect to other electronic devices. 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.

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. Web Phishing is a way of such fraud, which uses social engineering technique through short messages, emails, and WeChat 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. 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 . 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 . 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.

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 and white list, detection based on Uniform Resource Locator (URL) features, detection based on web content, and detection based on machine learning. The ant phishing 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.

A. SCOPE

The scope of any appraisal should include the following:

-Provides Provide security to web site by detecting phishing website on Web.

-Identify sevral types of attack on payment, online shopping and piracy copyies of web site.

-Improve working of cybercrime laws by helping ai based detection of phishing sites

-Increase commitment to organizational goals, develop firewall toward hacking sites.

II. PROJECT OBJECTIVES

When we judge whether a specific website is web phishing, the direct way is to use a white list or black list. We may search the URL in some database and decide. There are two ways to find phishing web site . The first way includes five heuristics to enumerate simple combinations of known phishing sites to discover new phishing URLs. The second way consists of an approximate matching algorithm that dissects a URL into multiple components that are matched individually against entries in the blacklist. Many well-known browser vendors such as Firefox and Chrome also used a self-built or third-party black-white list, to identify whether the URL is a phishing site. This method is very accurate, but its blacklist or whitelist usually relies on manual maintaining and reviewing. Obviously, these methods are not real time and may cost a lot of time and effort. Another phishing detection way is to analyze the features of URL. For example, sometimes a URL looks similar to the famous site URL or contains some special characters in the URL. used one concept of intra URL relatedness and evaluate it using features extracted from words that compose a URL based on query data from Google and Yahoo search engines. These features are then used in machine-learning-based classification to detect phishing URLs from a real data set. This method is efficient and economical because it utilizes the preexisting knowledge of the URL, which has a fast detection speed and a lower cost. However, we cannot fully exploit the characteristics of phishing in terms of an URL only because the essence of the scheme is to fraud by means of web content. Phishing attackers are very likely familiar with URLs and easily tailor their URLs to avoid detection; therefore this method will result in a lower detection rate if only the information of the URL is checked. 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. proposed an approach to detect phishing web page using Earth mover's distance (EMD) to measure web page visual similarity. The accuracy rate of this method is high. But at the same time the downside is a need to collect large amounts of data as a priori knowledge. With the popularity of machine learning, phishing detection has focused on the use of machine learning algorithms. This method integrates URL text features, domain name features, and web content features into a unified detection basis. presented a machine learning algorithm based on phishing detection using only lexical and described an approach to classifying URLs automatically as either malicious or benign based on supervised learning across both lexical and host based features. In general, the essence of these methods of machine learning detection is to map all the features of the phishing website into the same space and then to use the machine learning and data mining algorithms to detect phishing.

III. PROPOSED SYSYEM

To remove all the disadvantages of conventional methods, a system is proposed which is helpful for phishing attack detection.

An website based portal to tracking list of phishing site for safe our system's from hacking and various kind of attacks perform for collecting user confidential information.

This system is used to computerize all these activities.

There are two kinds of users: 1. Admin 2.User

1. Admin has access for our web servers to add list of phishing site that collected from various sources. This data provided to ai algorithm for avoid such web site and site having same nature.

2. User has access of one address bar in which user browse sites and if can not found as phishing site then auto open in next tab

A. SYSYEM DESIGN

A data flow diagram (DFD) is use a very small number of primitive symbols to represent the functionality performed by the project and the flow data among the different functions of the project.

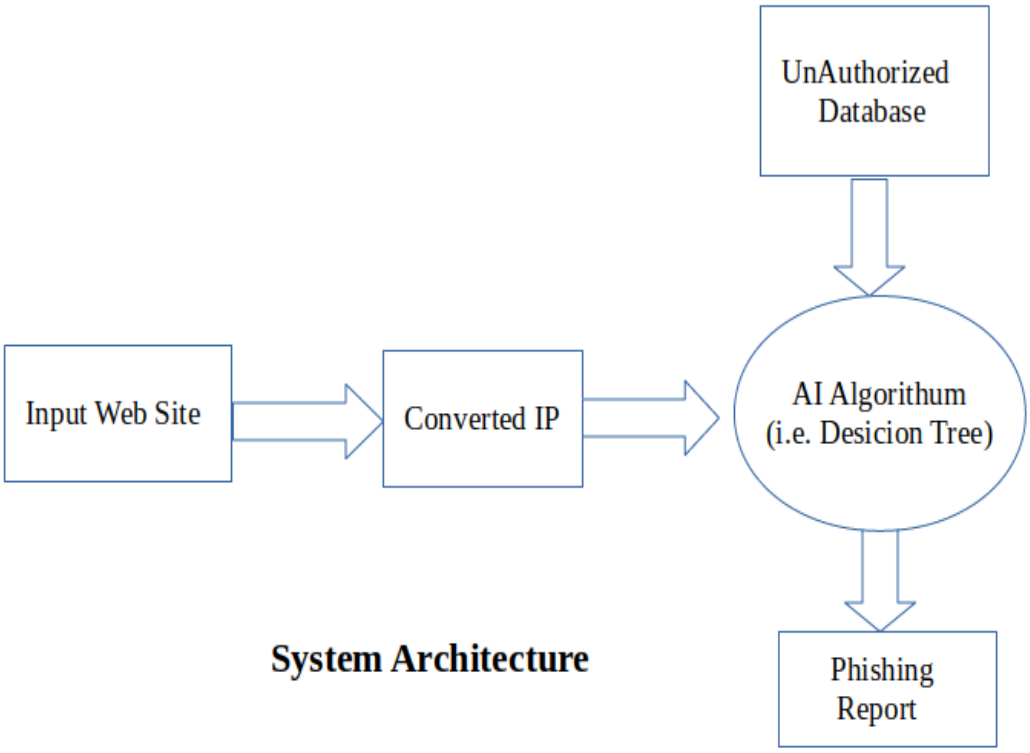

Fig.1 System Architecture

IV. IMPLEMENTATION

Classification can be defined as an estimate of a particular outcome, based on specific qualifications, starting from the training data. To estimate the results, a particular classification algorithm works on a set of qualifications and a training set containing the relevant result, often referred to as the target or estimated quality. The algorithm tries to predict the results and investigate possible relationships between qualifications. Then, the algorithm is given an unseen data set, called the set of estimates, containing the same set of attributes, with the exception of an unknown set of estimates. The algorithm analyses the input and generates an estimate. Prediction correctness indicates that the algorithm used is "good".

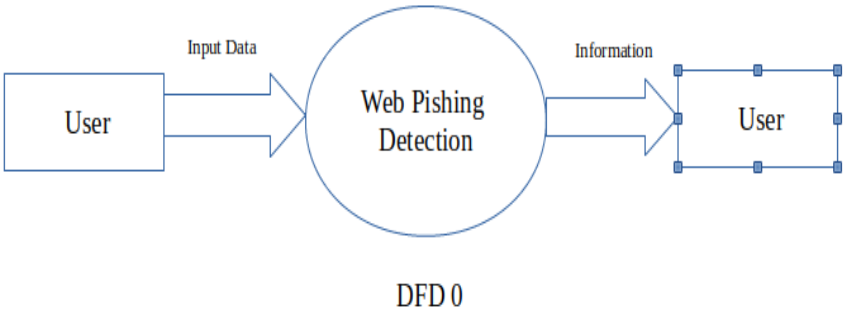

Fig.2 DFD0

After the preliminary stages of the data mining process, the choice of parameters and the choice of data set to be tested will affect the performance of the model that will be visible in the applications. For this reason, the result of the comparison will depend on the chosen classification algorithm. The data set (Table 1) used in this study is composed of the words that have the most used spam mail feature today. Each word has its own weight. These weights have been given a higher weight on the words that will cause the person to feel excited, fearful and hateful. In addition, these words are created with data mining, existing harmful words and harmful content created by the sites. With the Bayesian classifier-like algorithm, the weights of the words are calculated and a spam word count is made. At the same time certain rules are applied to prevent phishing attacks. Firstly, phishing attack links are detected on the Internet. In

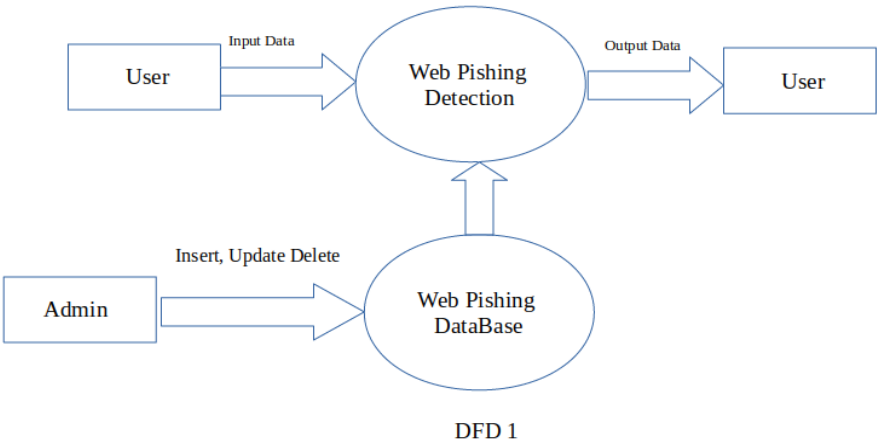

Fig.3 DFD1

This approach works efficiently in large dataset. This also removes drawback of existing approach and able to detect zero day attack .Machine Learning based classifiers are efficient classifiers which achieved accuracy more than 99%. Performance depends on size of training data, feature set, and type of classifier. Limitation of this is it fails to detect when attacker use compromised domain for hosting their site.

V. CONCLUSION

In this paper, we analyze the features of phishing websites and present two types of feature for web phishing detection. In this paper, we have offered an intelligent system for detecting phishing websites. The system acts as an additional functionality to an internet browser as web page based extension that automatically notifies the user when it detects a phishing website. The system is based on a machine learning method, particularly supervised learning.

REFERENCES

[1] R. Kiruthiga, D. Akila,International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-2S11, September 2019

[2] The Anti-Phishing Working Group. "What is Phishing?"URL:http://www.antiphishing.org/ (March 2019)

[3] Author Unknown. "How to Obscure any URL". PC-Help. January 2002. URL: http://www.pc- Help.org/obscure.html (March 2004)

[4] H.-C. Huang, Z.-K. Zhang, H.-W. Cheng, and S. W. Shieh, "Web application security: Threats, countermeasures, and pitfalls,"The Computer Journal, vol. 50, no. 6, pp. 81–85, 2017.

[5] https://en.wikipedia.org/wiki/WeChat.

[6] K. Rekouche, Early phishing, 2011.

[7] http://www.antiphishing.org/.

[8] Microsoft, "20% Indians are victims of online phishing attacks: Microsoft,"IANS,2014, http://news.biharprabha.com/.

[9] Chaitanya Bhandari Author of Cyber Surksha ith The Information Technology Act 2000, India.