Proceedings of the 14th INDIACom; INDIACom-2020; IEEE Conference ID: 49435
2020 7th International Conference on "Computing for Sustainable Global Development", 12th - 14th March, 2020
Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), New Delhi (INDIA)

Detection of Maliciious Content using AI

Yogesh Pingle
Department of Information Technology
Vidyavardhini College of Engineering
Vasai, INDIA
yogesh.pingle@vcet.edu.in

Sushmita Patil
Department of Information Technology
Vidyavardhini College of Engineering
Vasai, INDIA
sushmita.patil@gmail.com

Sneha N.Bhatkar
Department of Information Technology
Vidyavardhini College of Engineering
Vasai, INDIA
sneha24bhatkar@gmail.com

Shruti Patil
Department of Information Technology
Vidyavardhini College of Engineering
Vasai, INDIA
1999shrutipatil@gmail.com

***Abstract*—Most disruptive action which is performed on the Internet is phishing. Personal files or any business-related information will be at risk if a user gets attack by such actions. These attacks are getting increased day by day. Some attack is carried by inducing a URL which looks similar to a legitimate URL to steal the user's important files. Aim of the project is to detect malicious sites made by attackers to steal user's personal information in the aim of conducting illicit activities. Features from the submitted URL will be extracted. Then decision tree algorithm will use these features and will classify the site as malicious or genuine.**

Keywords— Information Gain (IG), Support vector machine (SVM), Uniform resource locator (URL) and Iterative Dichotomiser 3 (ID3)

I. INTRODUCTION

In recent years, Internet had an enormous growth and there has been also enormous growth of web service. Even web attacks have increased in large numbers and even improved in quality. One of the popular attacks which is growing since many years is Phishing. One of the malicious attacks, phishing is carried out to steal user's personal and important information such as bank details, passwords and other important files which may cause harm to user if used for illicit activities. Phishing is done with different communication forms such as instant messaging, email, SMS, etc. But mostly users get tricked by phishing attack is caused through uniform resource locator (URL) [3].

Business are the most prone to get attacked because if a user is tricked to access a malicious URL, it is easy way to access the user's information so that the attack can gain access to the business network. The RSA 2012 annual fraud report states that since 2011, 59% phishing attacks has increased in 2012. Loss of $1.5 billion has been estimated as the losses in 2012 globally. It has been estimated by 2013, the losses will be difficult to handle for the businesses.

Individual Internet users can also be prone to such attacks. .COM namespace, a top-level domain contained most unique domain names which is used for phishing website also having most numbers of attacks. This information is not good news, as many Internet users may have various accounts on different websites related to social media, banking, emails etc. Users doesn't know that this information stored are not safe on internet. The malicious links are placed on those genuine web pages which users unknowingly access it. Like genuine web pages, malicious web pages are created which take users to the attacker' server instead of genuine web server. Malicious or phishing URL have unique characteristics which are different from the genuine URLs.

Attacker attacks through the URL, which the user is not able to identify it.

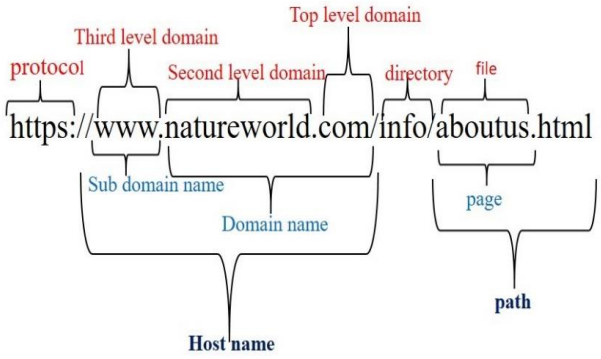

Fig. 1. Structure of URL

The above figure shows an example where URL structure for a website is defined. Protocol is used to access the page. Domain name is used to identify the server which owns that particular webpage. Then there is Sub domain, also called as registered domain name. There is suffix which is called top domain name. The Domain name part has to be registered under domain name Register. The part containing host name has sub domain name and a domain name. Attacker can easily

 ISSN 0973-7529; ISBN 978-93-80544-36-6