IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 5, Issue 3, March 2016*

# Visual and Textual Content based Anti-Phishing

**Mr.Digambar Pawar[1], Mr.Anuraj Jagdale[2], Mr.Rohit Hinge[3], Mr.Rupesh Gangtire[4]**

BE, Computer, PES's College of Engineering, Phaltan, India[1,2,3,4]

**Abstract**: Increasing the number of internet usage over the world, rapidly increases the number of phishing attacks and attacker acquires credential information of user. There are various techniques available to defend phishing attacks such as Microsoft blacklisted database and Google blacklisted database it contains only list of phishing websites URL which is previously detected. The study is to search on the internet utility which is mainly used for detecting phishing attacks and monitoring the authentication through pop-up's on every fraud web site and use images containing institution's corporate logos and artwork. Using such type of internet utility we can prevent the phisher's which use such criteria to fool the user and cause the economical disturbance or waste the money, using such types of tools like CANTINA we can secure our system from the fraud people and we can use the internet facility free without fear of phishers.

**Keywords**: URL (Uniform Resource Locator), CANTINA(Carnegie Mellon Anti-Phishing and Network Analysis Tool).

## I. INTRODUCTION

Phishing is criminal activity design to steal customer valuable data such as user-name, password to get a authorized user identity to theft, worms or viruses install key loggers on user computers is also referred as "phishing". Using such types of tricks or dishonesty phisher can deceive the users and get the benefits using the credit cards number, bank account numbers and social security numbers. Today, the most common type of phishing is E-mail Phishing. In typical scenario, a phisher sends fraudulent E-mail which having a fraud link, user do not see the received link carefully and user clicks on that link, after visiting that link user will redirect to phish web site. Its look likes as original but actually it is fraudulent created by phisher. User may trust on that site because it is created by expert judgment and there may be chances to enters the credential information on that fraudulent web site. So phisher easily gets the credentials information and get achance to unauthorized access of user accounts. Following are some of the classical phishing techniques used by phisher:

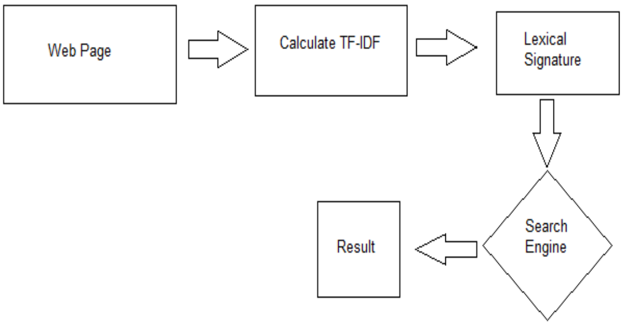
- **Deceiving a user into believing a message comes from a trusted source:** Using such types of techniques the phishers shows the message to the user are come from trusted site.
- **Deceiving a user into believing that a web site is a trusted institution:** Using such a type of a technique the phishers or the hacker show or design their websites in such way that user must trust on that websites.
- **Deceiving a spam filter to classify a phishing email is legitimate:** In today's life E-mail is one of main techniques used by the phisher's where he send fraudulent link. When user click on that link user redirect to phish web site.

To avoid such type of attacks the user can be use different techniques. The commonly used technique is content based anti-phishing by using TF-IDF. Another used technique is visual similarities by using EMD.

## II. LITERATURE SURVEY

### (A) Content Based Anti-Phishing:

Mr. Zhang [2] developed a content based approach known as Carnegie Mellon Anti-Phishing and Network Analysis Tool (CANTINA), for anti-phishing by employing the idea of robust hyperlinks. In this method first calculate the TF-IDF of each web page. TF-IDF is an algorithm usually used for information retrieval and generates a lexical signature by selecting a few terms. Signature supplies to search engines for example Google and then matches the domain name of current web page and several top search result to evaluate a current web page is legitimate or not.



The Term Frequency is simply the number of times that term occurs in specific document. The Inverse Document Frequency measure the importance of term. In CANTINA, Lexical Signature generated on few unique terms then that signature is applied to search engine. The Lexical Signature for given web pages is matches with billons online web Pages. The classification is based on measurement from the page rank. Thus, based on statistical information from the attack history data page is classified into legitimate or phish. Other content based technique Bayesian anti-phishing toolbar(B-APT) is designed to identify phishing web sites by using open source by sun filter on the bases of tokens which are extracted by a Document Object Module (DOM) analyser.

**Advantages:**

- This approach detects phishing by Lexical Signature.

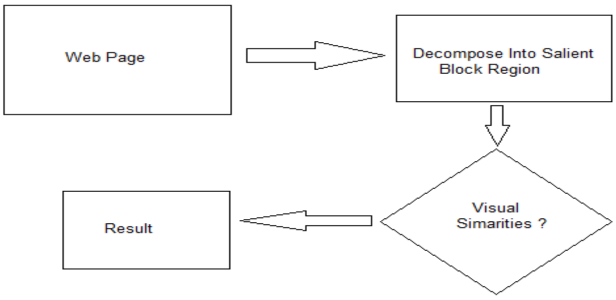
DOI 10.17148/IJARCCE.2016.53236

**Disadvantages:**

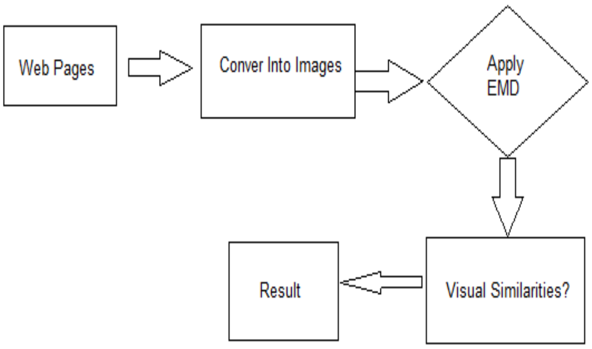
- This approach cannot detect the phishing at dynamically.
- This approach requires the use of historical data for that page.

### (B) Visual Based Anti-Phishing:

The concept of visual approach to phishing detection was first introduced by Mr. Liu [1] This approach which is oriented by a DOM-based visual similarity of web pages , first decomposed the web pages into salient block regions.



The visual similarity between two web pages is then evaluated by three matrices, namely, block level similarity, layout similarity and overall style similarity, which are based on matching on salient blocked region. They first converted HTML web pages into image's and then employed the Earth Movers Distance (EMD) by Carlo Tomasi[3] method to calculate the similarities of the images.



**Advantages:**

- This approach detects phishing by using EMD method.

**Disadvantages:**

- This approach is only investigates phishing detection at the pixel level of web pages without considering the text level.

This technique involves choosing a number of attracted regions and associating each region with the attractor that has the largest attraction to it. The MARS uses the five attractors, one for each corner of the image called background attractor and one for the center of image called object attractor. This consists of the fact that their database contains images of single object. The shape of the boundary of the extracted object is represented by means of Fourier Descriptors (FD). Boolean operators are used to formulate the complex queries. The desired features of image can be specified by pointing an image database that has such property or by directly choosing the color from given palette and texture from available set of pattern in the database.

For Visual Based Anti-Phishing, we use image Classifier. First we retrieve the web page from web then we generate the signature by calculating EMD. Signature has two components as Feature and corresponding weight.

The EMD is adopted to find out the distance or dissimilarity between two web pages. Suppose we have two web pages Sa & Sb where (Sa = m feature units and Sb = n feature units) Then we calculate the Distance Matrix D.

### (C) Microsoft and Google Blacklisted databases:

The most popular and widely-deployed techniques, are based on the use of blacklists of phishing domains that the browser refused to visit. For Ex, Microsoft has recently integrated a blacklist-based anti-phishing solutions into its Explore(IE)7 browser. The browser queries lists of blacklisted and white listed domains from Microsoft servers and makes sure that the user is not accessing any phishing sites. Microsoft solution is also known to use some heuristics to detect phishing symptoms in web pages. Obviously, to date the company has not related any detailed public information on how its anti-phishing techniques function. Other browser integrated anti-phishing tool includes Google Safe Browsing, Net Craft tool bar, eBay tool bar and McAfee site Advisor.

**Advantages:**

- It detect website based on their own blacklisted database.

## III. CONCLUSION

Visual based and content based anti-phishing both are widely used anti-phishing techniques. In visual based technique the site is detected by comparing the similarities between the web pages and it detect site is phish or not. In content based anti-phishing the phish web sites are detected by using the lexical signature. Also other technique is blacklisted databases which contain the sites URL which are already detected as a phish site.

## REFERENCES

[1] Haijun Zhang, Gang Liu, Tommy W. S. Chow, Senior Member, IEEE, and Wenyin Liu, Senior Member, IEEE, "Textual and Visual Content-Based Anti-Phishing: A Bayesian Approach," IEEE TRANSACTIONS ON NEURALNETWORKS, VOL. 22, NO. 10, OCTOBER 2011.

[2] Y. Zhang, J. Hong, and L. Cranor, "CANTINA: A content-based approach to detecting phishing web sites," in Proc. 16th Int. Conf. World Wide Web, Banff, AB, Canada, May 2007, pp. 639–648.

[3] A. Y. Fu, W. Liu, and X. Deng, "Detecting phishing web pages with visual similarity assessment based on earth mover's distance (EMD),"IEEE Trans. Depend. Secure Compute, vol. 3, no. 4, pp. 301–311, Oct.–Dec. 2006.

[4] Colin Whittaker, Brian Ryner, Marria Nazif, "Large-Scale Automatic Classification of Phishing Pages"

DOI 10.17148/IJARCCE.2016.53236