Phishing Webpage Detection for Secure Online Transactions

Sathish .S, Thirunavukarasu .A

Department of Computer Science and Engineering (PG), Regional Centre, Anna University, Madurai - 625 002.

Abstract

Problem Statement: Phishing Websites are duplicate Web pages created to mimic real Websites in-order to deceive people to get their personal information. Because of the adaptability of their tactics with little cost Detecting and identifying Phishing Websites is really a complex and dynamic problem. In this project, a novel approach is proposed for detecting phishing Websites. In the proposed model three layers of criteria are used. In the first layer Google page rank and IP address in URL are used. In the second layer domain name characteristics are used. In the third layer quality of the web page content are used. Fuzzy logic is used to classify the web pages according to their rank.

Keywords

Google Page Rank, Domain Name, Fuzzy Logic

I. Introduction

Phishing can be defined as a criminal activity using social engineering techniques. The most common example of phishing is an email asking to enter account/credit card information for Electronics Commerce websites (eBay, amazon etc.) and online banking. It is a kind of merger of Web technology and social engineering. The most popular phishing scams are carried out using phishing web pages. Phishing web pages are forged to mimic certain genuine companies' web pages. It is also kind of attacks in information security. Phishing, vishing and smishing are all ways for a thief to use current technology to get customer personal account information to use for fraudulent purposes.

In 1965 fuzzy logic was introduced by Zadeh as modification of classical set theory. Fuzzy set theory enables the processing of imprecise information by means of membership function. Fuzzy logic allows intermediate categories between notations such as true/false, hot/cold, black/white etc. as used in Boolean logic. In fuzzy system, values are indicated by a number in the range of 0 to 1 where 0 represents absolute falseness and 1 represents absolute truthfulness.

μ_A : X → [0, 1], where

$\mu_A(x) = 1$ if x is totally in A;
$\mu_A(x) = 0$ if x is not in A;
$0 < \mu_A(x) < 1$ if x is partly in A.

Mamdani's fuzzy inference method is the most commonly seen fuzzy methodology. There are primary GUI tools for building, editing, and observing fuzzy inference systems in the toolbox like Fuzzy Inference System (FIS) Editor, Membership Function Editor, Rule Editor, Rule Viewer and Surface Viewer.

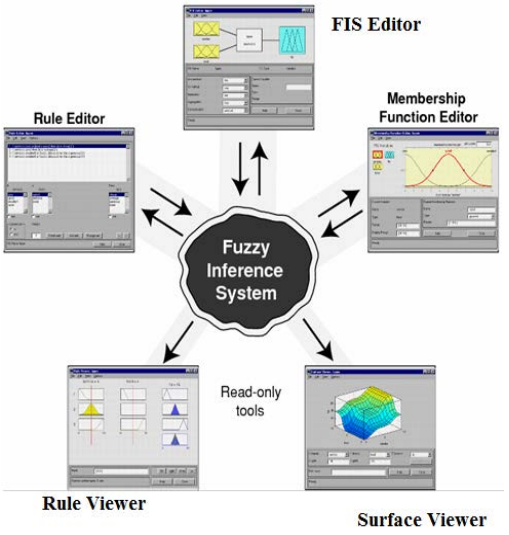

Fig. 1 Fuzzy Inference Process Diagram

The FIS Editor handles the high-level issues for the system, The Membership Function Editor is used to define the shapes of all the membership functions associated with each variable. The Rule Editor is for editing the list of rules that defines the behaviour of the system. The Rule Viewer and the Surface Viewer are used for looking at, as opposed to editing, the FIS. The Rule Viewer is a MATLAB® technical computing environment based display of the fuzzy inference diagram shown in Fig 1. The Surface Viewer is used to display the dependency of one of the outputs on any one or two of the inputs that is, it generates and plots an output surface map for the system.

II. related work

Phishing attacks are becoming more frequent and sophisticated. Phishing scams have become a problem for

Manuscript received March 5, 2015
Manuscript revised March 20, 2015

online banking and e-commerce users. In the Literature several phishing detection scheme has been analyzed.

In [1], phishing Web page detection using Earth Mover's Distance (EMD) has been proposed to measure Web page visual similarity.They used EMD to calculate the signature distances of the images of the Web pages. The approach works at the pixel level of Web pages rather than at the text level. Bessedik Imene and Taghezout Noria have proposed multi-agent approach provides a practical way to implement a Web-based DSS. They have used an efficient tool that helps users find information resources available as an online service within Intranet. The decision-making is not only guided by the information provided by DSS but rather than the Web technology, the process is entirely based on communication between ISP Agents and Web agent.

In [2], a heuristic method is proposed to determine whether a webpage is a legitimate or a phishing page. This scheme could detect new phishing pages which black list based anti-phishing tools could not. they first converted a web page into 12 features which are well selected based on the existing normal and fishing pages. A training set of web pages including normal and fishing pages are then input for a support vector machine to do training. A testing set is finally fed into the trained model to do the testing. Compared to the existing methods, the experimental results show that the proposed phishing detector can achieve the high accuracy rate with relatively low false positive and low false negative rates.

In [3], a new approach called multi-tier classification model for phishing email filtering nave been proposed. They have proposed an innovative method for extracting the features of phishing email based on weighting of message content and message header and select the features according to priority ranking. Their method will also examine the impact of rescheduling the classifier algorithms in a multi-tier classification process to find out the optimum scheduling.

In [4], new approach based on a Neuro-Fuzzy scheme to detect phishing websites and protect the customers performing online transaction. Hybrid Neuro-Fuzzy technique was used to developed the proposed detection and protection scheme that offered an effective technique. Using 2-fold cross-validation, the overall results demonstrates that the proposed Neuro-Fuzzy systems with five inputs offers a higher accuracy and can be effective in detecting phishing sites with a high accuracy in real-time. The proposed system introduced few inputs (Legitimate site rules, User-behavior profile, Phish Tank, User-specific sites, Pop-Ups from emails). The idea is to utilize a Neuro-Fuzzy Scheme with 5 inputs to detect phishing sites with high accuracy in real-time. 2-Fold cross-validation is applied for training and testing the proposed model. A total of 288 features with 5 inputs were used.

They have proposed a new way of phishing page detection based on Transductive Support Vector Machine (TSVM) that is independent of attack method and not change the users' behaviour. Extract the features of webpage image and sensitive information in page, which they could reflect the web page properties completely. Then the phishing pages are classified by TSVM algorithm. The experiments show that the proposed method performs well and improve the accuracy and precision, also have more completely. The results are only a preliminary investigation of detecting phishing web page using TSVM, and there are much can be done to improve the performance, such as the classifier of TSVM should be more flexible, which makes TSVM learn date more efficiently.

In [5], a phishing detection approach based on checking the webpage source code is proposed. Common properties of phishing attack in the WebPages are Request URL (RURL), URL of Anchor (AURL), Server Form Handler (SFH), Logos, Suspicious URLs, User input, short lived and lack of familiarity with English and they have considered 8 phishing characters like different domain, external link for images, suspicious URLs, domain tag, iframe, email, suspicious script and popup window. They extract some phishing characteristics out of the WWW Consortium standards to evaluate the security of the websites, and check each character in the webpage source code, if they find a phishing character; they will decrease from the initial secure weight. Finally they 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

III.PROPOSED SYSTEM

The proposed phishing website detection system consists of three stages Such as data collection, fuzzy rule base and classification. For this phase only data collection part has been implemented and it will be explained thoroughly in the following section. The remaining two module work will be given briefly. They will be elaborated in the next phase of the project.

1. Data Collection
2. Fuzzy Rule Base
3. Classification

A. DATA COLLECTION

The data collection part consists of finding phishing and genuine web pages and analysis. In total 38 URLs have been identified. The factors used to analyse the URLs are

google page rank, IP number in URL, whether host name present in the URLs, whether the URLs contain word that is similar to host name but not same, whether there are long unwanted extra words present in the URL after or before the host name, certificate, Lots of spelling mistakes in the webpage contents and the contents been copied from other popular web pages. The web pages collected are given in appendix.

A.1 Page Rank

Google page rank is one of the reliable information to identify whether one web page is genuine or not. Google ranks every web pages listed in it. Most visited and trusted web pages will be given high ranks. Least visited or unknown web pages will be given very low rank, sometimes zero rank.

For example from the collected URLs Page rank of the web page

http://amazon.com

http://www.alliancebhd.net/ibank/en&source=hp&biw=1280&bih=679&q=a+man/contact.asp.htm is not available this means the former web page can be trusted and the later one is suspicious.

A.2 IP Address in URL

Any major well known company or organization would like to have their name in their URL. If IP address was present in any URL that claims to be genuine one it could be a phishing web page.

For example. The URL given below claim that it is from hotmail but the IP address present in the web page make sure that it is a fake one.

http://91.239.233.31/~snkbiz24/hotmail/

A.3 Domain Name in URL

Phishing web pages can be identified based on how the host name is used in the domain. Some have same domain name of genuine site but with some extra unwanted characters. Some websites have host name that look similar to host name. Some do not have host name at all. The following examples demonstrate them. The first URL has domain name but it has some extra words with it so it may be a suspicious one. The second one has a domain name which is similar to genuine one but not same so it may be a fraudulent one.

http://amzon.cos54897987654.altmfg.com/755961a93b4c0f7ad08688225392cbf1/index/webscr.php

http://verify.account.information.secuirty.amazowebs.com/6273556f86e7a25b93e4ff6100101ec3/

A.4 Quality of the Web Page

The quality of the content of the web page concerned is another factor used to decide its genuineness. If the image present is not at good quality or the words in the web page contain lots of spelling mistakes or if the overall content was copied from another web page it can be concluded that the web page is fake one.

B.Fuzzy Rule Base

Fuzzy rules are written to help the inference system to make a logical conclusion about the genuineness of the web page presented. The rule base consists of three input parameters and one output parameters; it contains all the "IF-THEN" rules of the system. For each entry of the rule base 1, each component is assumed to be one of three values and each criterion has three components. Therefore, the Rule base 1 contains (33) = 27 entries. The output of Rule base 1 is one of the e-banking phishing website rate fuzzy sets (Genuine, Doubtful or Fraud) representing URL &Domain Identity criteria phishing risk rate. The system structure for URL & Domain Identity criteria is the joining of its three components (Abnormal URL of Anchor, Abnormal DNS record and Abnormal URL), which produces the URL & Domain Identity criteria (layer 1).

In layer 2, there are three input parameters and one output parameter. The rule base contains (33) = 27 entries, and the output of Rule base for layer 2 is one of the e-banking phishing website rate, which consider as a fuzzy sets like Legal, Uncertain or Fake.

In layer 3, there are four input parameters and one output parameter. The rule base contains (34) = 81 entries, and the output of Rule base for layer 3 is one of the e-banking phishing website rate, which consider as a fuzzy sets like Legal, Uncertain or Fake.

B.1 The rule base for integrating Layer1, Layer2 and Layer3

The system structure is consists of three input parameters and one output parameter. Input parameters are layer 1, layer 2, and layer 3. Output parameter is website risk. The structure and the entries of the rule base for e-banking phishing website rate are shown in Table 5.4. The rule base contains (33) = 27 entries, and the output of final phishing website rate, which consider as a fuzzy sets like Legitimate, Suspicious, Phishy, Very phishy.

C. Classification

In classification part the FUZZY LOGIC , SVM or WEKA mining tool can be used to classify the web pages based on the risk factor assigned to them. In the classification, we have to preprocess the data with the help of Weka data mining tool.

True Positive (TP): Which correspondence to the number of positive example correctly predicted by the classification model.

False Negative (FN): Which correspondence to the number of positive example wrongly predicted as negative by the classification model.

False Positive (FP): Which correspondence to the number of negative example wrongly predicted as positive by the classification model.

True Negative (TN): Which correspondence to the number of negative example correctly predicted by the classification model.

The counts in a confusion matrix can also be expressed in terms of percentages.

The True Positive Rate (TPR) or sensitivity is defined as the fraction of positive examples predicated correctly by the model.

The True Negative Rate (TNR) or specificity is defined as the fraction of negative examples predicated correctly by the model.

The False Positive Rate (FPR) is the fraction of negative examples predicated as a positive class.

Finally**, the False Negative Rate (FNR)** is the fraction of positive examples predicated as a negative class.

IV. Implementation & Results

The classification is done using the weka data mining tool. The results for classification and integration of three layers are analysed and the results are as follows,

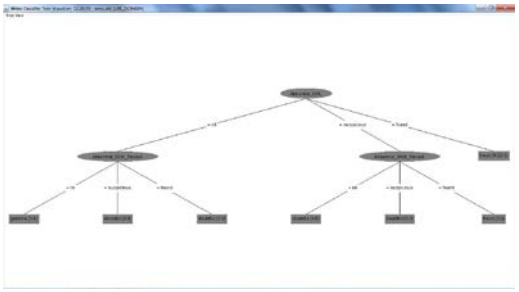

Fig. 2 Domain Identity Risk

Decision tree of URL &Domain Identity risk (Layer 1)

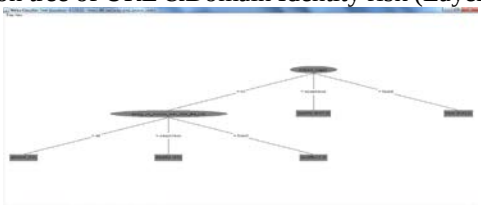

Fig. 3 Source Code Risk

Decision tree of Security & Source code risk (Layer 2)

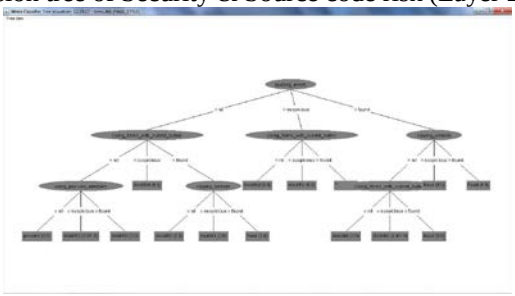

Fig. 4 Page Style & Content Risk

Decision tree of Page Style & Contents Risk (Layer 3)

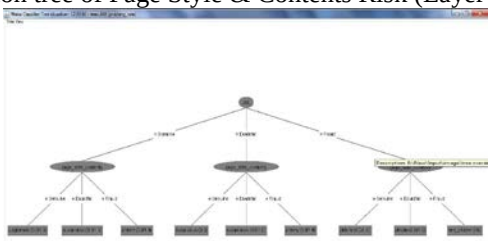

Fig. 5 Phishing Website Risk

Decision tree of Overall phishing website risk (Integrating Layer 1, Layer 2 and Layer 3)

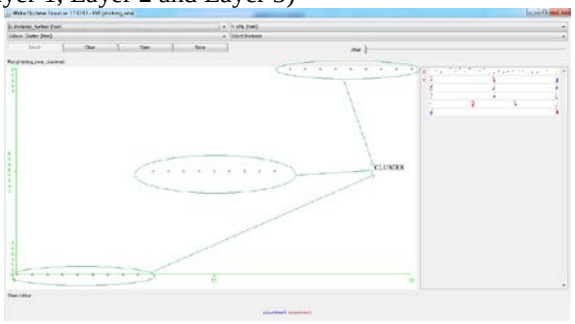

Fig.6 Weka Tool

Clustering in WEKA tool for Phishing Website Risk(Integrating Layer 1, Layer 2 and Layer 3)

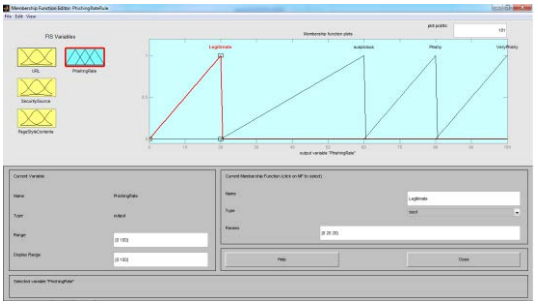

Fig. 7 Membership Function Editor

Membership Function Editor (Integrating Layer 1, Layer 2 and Layer 3)

The data collection part consist of finding phishing and genuine web pages and analysis. In total 38 URLs have been identified. The factors used to analyse the URLs are google page rank, IP number in URL, whether host name present in the URLs, whether the URLs contain word that is similar to host name but not same, whether there are long unwanted extra words present in the URL after or before the host name, certificate, Lots of spelling mistakes in the webpage contents and the contents been copied from other popular web pages. The web pages collected are given in appendix.

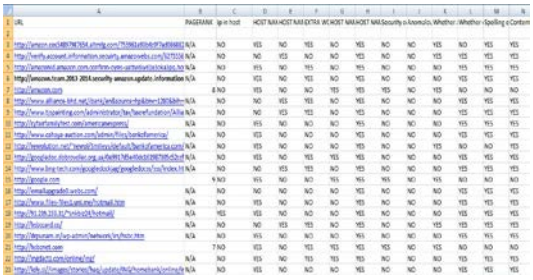

Fig. 8 Data Sample

V. Conclusion

Most popular websites of today such as eBay, amazon, ICICI bank, Facebook etc., provide the strongest of fraud detection techniques. Still the intruders find a way so that they gain an entry to the system using fake websites. Though a number of attempts have already been done in this area, this project provides much more capability in the identification of phishing websites using a number of parameters. Classification tool WEKA is used for the classification of web sites using the determined parameters. Fuzzy inference system is employed to identify and report the end user of the risk factors. In the future, more efficient classification tools with much better classification algorithms can be used to further refine the classification based on the risk characteristics.

References

[1] Anthony Y. Fu, Liu Wenyin and Xiaotie Deng, "Detecting Phishing Web Pages with Visual Similarity Assessment Based on Earth Mover's Distance" , IEEE Transactions On Dependable And Secure Computing , Vol. 3, No. 4, pp.301-311 December 2006.

[2] Mingxing He , Shi-Jinn Horng ,Pingzhi Fan,Muhammad Khurram Khan ,An efficient phishing webpage detector , Elsevier, Expert Systems with Applications 38 (2011) 12018–12027.

[3] Rafiqul Islam , JemalAbawajy , A multi-tier phishing detection and filtering approach, Elsevier, Journal of Network and Computer Applications 36 (2013) 324–335.

[4] Maher Aburrous, Hossain.M.A , KeshavDahal , FadiThabtah, "Intelligent phishing detection system for e-banking using fuzzy data mining", Elsevier-Expert Systems with Applications, Vol. 3, No.7, pp.7913–7921, 2010.

[5] Mona GhotaishAlkhozae, Omar Abdullah Batarfi, "Phishing Websites Detection based on Phishing Characteristics in the Webpage Source Code", International Journal of Information and Communication Technology Research, Vol.1, No.6, pp.284-291, October 2011.