PHISHSTORM: Detection Mechanism for Finding the Phishing and Legitimate Websites by Implementing the Search Engine

R.Murugeswari , G.Elizabethrani, G. Rama Sakthi, B. Atchaiya

Abstract*: This paper proposes software for detecting the phishing and legitimate websites. In that software, we build and explain the mechanism behind the detection of the phishing and legitimate websites. Phishing is defined as a fraudulent attempt, used to obtain confidential information through which an attacker emulates an actual person or institution by disguising them as an official person or entity through e-mail or other communication mediums. We created this project to safeguard the users from the phishing websites so that they won't lose their personal details like usernames, passwords and credit card details. Here, we combined Storm in the architecture, thereby detecting the phishing and legitimate URL's manually. A bloom-filter is a space-efficient data structure used to test the element either present in the set or may not be in the set. The simulation is carried out in the search engine, entering the search word, viewing the phishing and legitimate websites, downloading the files, give feedback about the websites and transformation is done wherever it is needed.

Keywords: Phishing websites, Storm Architecture, Bloom Filter.

I. INTRODUCTION

Phishing is currently one of the most lucrative cybercrime activities. Although accurately evaluating the financial loss caused by phishing is difficult, some surveys have been conducted, suggesting losses of several billion dollars every year.

In 2007, Gartner Research estimated a loss of $3.2 billion due to phishing scams [1]. Javelin Strategy & Fraud published a report that identity theft led to a loss of $54 billion in 2009, mostly due to cybercrime [2]. More recently, RSA estimated worldwide losses of $687 million due to phishing in the first half of 2012 [3].

Various techniques are used to perform phishing attacks, ranging from technical subterfuges (DNS cache poisoning, email spoofing, Web server takeover, etc.) to social engineering.

Phishing is a fraudulent attempt to obtain the sensitive information such as usernames, passwords and credit card details of the users by disguising as a trust worth entity in an electronic communication. Phishing is now considered as one of the cybercrime activities. The data, money and the credential information were stolen through fake websites and drive by download of malware etc., also come under the category of phishing. One common feature in the phishing is the obfuscated URL's to misdirect the users to fake websites or drive by downloads.

The Internet users are provoked to click on the rogue links that seems trust worthy is an easy task because of credulity and unawareness. To get relieved from this threat, the best method is to prevent the connection to phishing websites by the identification of phishing URL's. To infer intra-URL relatedness, the distributed real- time computation technique (Storm) is used. To check the availability of content or files related to the URL, the space-efficient data structures (Bloom filter) is used.

In this paper, we proposed an automated real-time URL phishing-ness rating system and combines it with the architecture to protect the users against phishing content: Phish Storm. The underlying method targets identification of phishing and legitimate URL's manually and they are based on the relativity of files or content that belongs to the URL. This paper explains the mechanism which is happened behind the detection of phishing websites. It explains how the websites are entered, how the websites are approved, on which basis they are classified as phishing or legitimate and overall mechanism is discussed here. Here the Bloom filter data structure is used. It is used to filter the URL's and classified them as good or bad URL.

The remainder of the paper is organized as follows. Section 2 discusses the related work about the detection of phishing websites. Section 3 presents the proposed work. The functions of the modules are proposed as steps in section 4. Finally, the experimental result is presented in section 5.

Revised Manuscript Received on December 05, 2019.
***** Correspondence Author
R.Murugeswari*, Computer Science and Engineering, Kalasalingam Academy of Research and Education, KrishnanKoil, India. Email: r.murugeswari@klu.ac.in
G.Elizabethranii, Computer Science and Engineering, Kalasalingam Academy of Research and Education, KrishnanKoil, India. Email: g.elizabeth@klu.ac.in
G.Ramasakthi., Computer Science and Engineering, Kalasalingam Academy of Research and Education, KrishnanKoil, India. Email: ramalakshmanan31@gmail.com
B.Atchaya,, Computer Science and Engineering, Kalasalingam Academy of Research and Education, KrishnanKoil, India. Email: batchaya@gmail.com

II. RELATED WORKS

The main aim of this paper is to detect the phishing websites using by implementing the search engine. Here there are some related papers which describes about the phishing website detection techniques.

In [7] proposed phishing websites using data mining. The main aim of this paper is to determine the effective set of features in terms of classification performance. The author was compared two known features and thereby selecting the least set of features of phishing detection. Experimental tests have been done using two features namely Information Gain and Correlation Features set methods. Moreover, two algorithms namely PART and IREP have been performed on the two different sets of selected features to distinguish the phishing and legitimate websites by the selection features.

In [8] described the comparative analysis of phishing and legitimate dataset using machine learning algorithm. It compares the machine algorithms such as Random Forest, C4.5, Decision Stump, Hoeffding tree, Rotation Forest, REP Tree and MLP and analyses which method provides the best results in the detection of phishing websites. The instances are named as 1 for "Legitimate", 0 for "Suspicious", and -1 for "Phishing". Experimental results show that Random Forest and REP Tree show the best performance on the dataset in the detection of phishing websites.

In [9] proposed ANN (Artificial Neural Network) multilayer perceptron algorithm for website classification with phishing characteristics. In this algorithm, a pattern is given to the input layer of the network and it is processed layer by layer until the output delivers the response. The value ranges from 0.1 to 1.0. If a minimum value is found, the error decrease and reaches the greater value than the existed value. The experimental result shows that the ANN-MLP classified 87.61% in the websites with and without features. It has an accuracy of 98.23% in the test phase. It provides better performance in the classification of websites with phishing characteristics.

In [10] proposed MCAR (Multi Class Classification Based on Association Rule) algorithm to detect the phishing websites. This algorithm method predicts the classification of phishing websites on the basis of spelling errors, personalization, long URL's prefix and suffix. This algorithm intersects the TID lists of frequent rule items of size. By using cardinality of the set to determines the frequent rule items. Experimental results show that MCAR has achieved an average 6.8%, 6.2% and 5.4% higher accuracy than SVM, CBA and NB algorithms.

In [11] compared different classifier ensemble approaches, i.e. random forest, gradient boosted machine, rotation forest, and extreme gradient boosting against single classifiers, i.e. decision tree, classification and regression tree, and creedal decision tree in the case of website phishing. The dataset is made up of 30 input features with 1 class label. The samples having 11,055 instances with the proportion belongs to positive (+1) class is 44.31% and negative (-1) class is 55.69%. Random Forest has the better performance among the classifier ensembles and C50 denotes the effective algorithms among the single classifier.

In [12] described e- mail extraction based anti-phishing technique. A neural network model is designed in order to detect the phishing websites through the e-mail. In this the e-mail is parsed firstly into a MIME, feature is extracted from it. Next, the features were established by the neural network and classified in to the corresponding categories.

In [13] examined the popularity and interest of students in Croatia with threats in the form of social engineering and phishing attacks. The author also presented for specific countermeasures at both user level and the organization level and a multilayer anti-phishing proposal.

In [14] proposed the Rough set algorithm to determine the reliability factor of phishing website. The reliability factor shows the originality of the website. By using this algorithm, to detect the distinctness of the parameters included in the phishing and shows the better analysis for detecting the originality of the website.

III. PROPOSED WORK

The proposed software is an efficient implementation pattern, that allows real time analytics using Big Data architectures like storm and advanced data structures based on Bloom filter. The storm is combined in the architecture part and protects the user against the phishing content. The undergone method shows the mechanism of detecting the phishing and legitimate websites manually. For that, we had created a search engine and if we enter the search word, then it will show the phishing and legitimate URL's related to the search word. The special feature is that all are done manually here. These works are done by the three important modules in our project and they are provider, admin and user. In this paper, we show the proposed feature sets the manual work method than the existed calculation technique.

- To infer intra URL relatedness, the automated real time URL computation technique storm is used in the architecture. By that, once the search word is entered, then the websites related to the search word will be shown automatically.
- To detect the phishing and legitimate URL's, the search engine is created. In search engine, after entering the search word, it shows the list of phishing and legitimate websites presented there.
- To check the availability of suitable or matched files related to the corresponding URL, the bloom filter data structure is used here.

IV. MODULE DESCRIPTION

This paper contains a homepage. The homepage contains three main modules.

The three main modules are

- Provider

- Admin
- User

A. *Provider:*

The provider is one who provides the websites. The provider is asked to register with a provider ID and a password. The provider ID was generated automatically. By using this provider ID and password, he/she have to login to the provider site.

In the provider side, he/she can do the following things:

- He/Shecan see only their registration details.
- He/She can upload the files in the corresponding websites.
- He/She can view the uploading files also.

B. *Admin:*

The admin is one who approves or rejects the websites. He/She is given the authority to approve/reject the websites, update the status of the websites whether it is phishing or legitimate. The admin has a username and password. By using this username and password, he/she can log in to the admin side.

In the admin side, he/she has an authority to do the following things:

- The admin has the rights to update the status of the provider. Once the provider has been registered, the admin has to either approve or reject the website. Or else we can't enter in to the provider site. After the admin has updated the status then only, we can enter in to the provider site and upload the files in the corresponding websites.
- The admin has to add the search word. Search word is nothing but what word we need to search related to the website or topic that he/she has to provide here.
- The search word ID was generated automatically. Using that ID, he/she can add the search word.
- The admin can add the related words. Here he/she can select the word ID; the corresponding search word will come. He/She can add the related words in the respective places.
- The admin can add the unwanted words. It is nothing but the words we don't need the meaning of the word such as what, where, when, this, that, is, are, was, were etc.,
- The admin can add the URL's. By selecting the search word ID, he/she can enter the URL and mention the URL as either good or bad. They can use one or more URL's for the same search word. They mention the URL status on the basis of previous assessment of the URL's.
- The admin can view the user's information.
- The admin can view all the files and check whether the files are related to the content or not and thereby updating the status of the files.
- The admin can view the related words with the corresponding search word.
- After the feedback is given by the user, the admin has to update the websites as either phishing or legitimate based on the user's feedback only.

C. *user*

The user is one who uses the websites. The user is asked to register with a username and a password. By using this username and password, the user can log in to the website. The user can do the following things:

- The user can search the data here. After log in, the user is asked to enter the search word. Once you have entered, then the relevant websites list will be shown with the classification (phishing or legitimate).
- The user can download the file and they can give the feedback to the admin.
- The feedback is given based on the four categories. They are as follows:
  - Good
  - Poor
  - Irrelevant info
  - Fake info
- The user can give any of the above four categories. Based on the maximum no of user's feedback on the above categories, the admin can update the URL's.
- Out of these four categories, good will be taken as legitimate URL and the remaining URL's (poor, irrelevant info, fake info) are considered as phishing URL's.

The supervised classification technique is used here to detect the phishing website. It tends to map an input into an output based on the example output pairs. In supervised classification, each pair consists of an input object and a desired output value.

Whenever the user clicks the URL, he can download the files in the URL and go through the content in the files, then he must give the feedback. The user gives the feedback by checking whether the content present in the files are relevant to the corresponding URL's given. After checking the relation, then the user gives the feedback on the choices of previously told four categories.

The URL's can be classified according to the below categories on the user's side.

1. If the user downloads the phishing website and give feedback as phishing website, then it is known as true positives.
2. If the user downloads the legitimate website and give feedback as phishing, then it is known as false positives.
3. If the user downloads the legitimate website and give feedback as legitimate, then it is known as true negatives.
4. If the user downloads the legitimate website and give feedback as phishing, then it is known as false negatives.

After collecting the feedback from the user, the admin watches and analyzes the

feedback clearly and update the status of the websites wherever the update is needed. The admin considers the feedback and gives priority to the maximum no of feedbacks given by the user to the URL, then the updating is performed, wherever it is needed.

V. EXPERIMENTAL RESULTS

The experimentation done for detecting the Phishing web site uses ASP.NET as front end and MySQL server as back end. In homepage, the browser submits a web form to the server and the server returns HTML page in response. It uses the HTTP commands to set a browser-to-server bilateral communication.

In admin page, it contains labels and textboxes for configuring and manipulating code to create HTML pages as it was created by using ASP.NET. MySQL is associated with web applications. MySQL server serves as back end. It deals with the coding part. The simulation is carried out in the search engine, entering the search word, viewing the phishing and legitimate websites, downloading the files, give feedback about the websites and transformation is done wherever it is needed.

This architecture explains the parts and concept about the detection of phishing websites. It is composed of internet, email server and phish storm. Phish storm contains the URL word extraction, features computation and prediction. When we search the websites though the internet connection, it will go through the Phish storm.

With the help of features in the phish storm, it detects the phishing websites and sends it to the email server and then it reaches the user. This is done with the help of ASP.NET Development server and the remaining is done.

A. SAMPLE SCREENSHOTS

Fig. 1. Homepage

Fig. 2. Admin Login

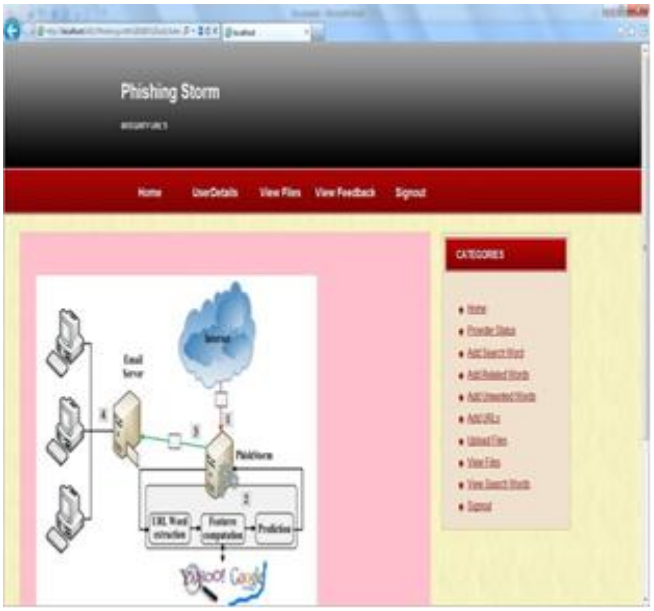

Fig. 3. Admin Home

Fig. 4. Add Search Word

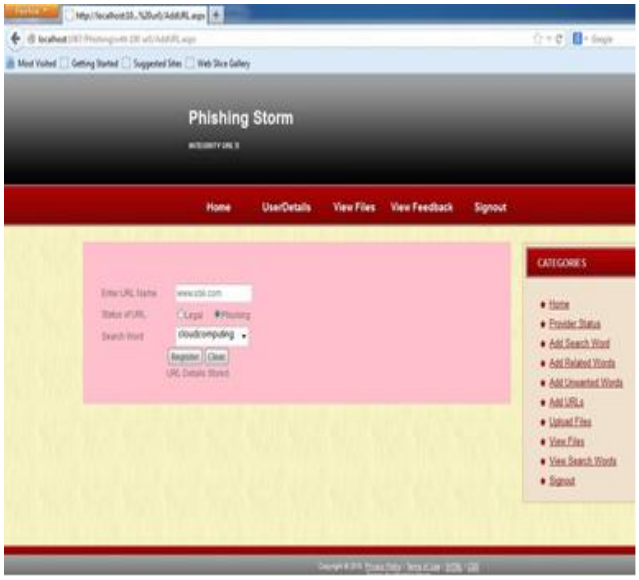

Fig. 5. URL Details

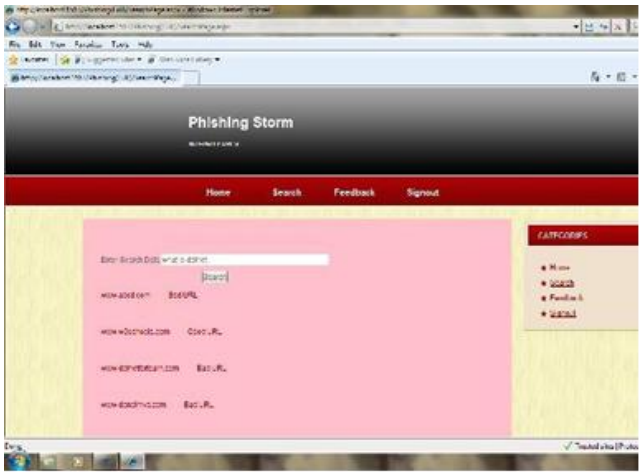

Fig. 6. Sample Output Screen

REFERENCES

1. "Gartner survey shows phishing attacks escalated in 2007," Gartner Research, Tech. Rep., 2007.
2. Javelin Strategy & Research shows the "Identity Fraud Survey Report 2010," Tech. Rep., 2010.
3. P. J. Nero, B. Wardman, H. Copes, and G. Warner, "Phishing: Crime that pays," in eCrime Researchers Summit (eCrime), 2011. IEEE, 2011, pp. 1–10.
4. APWG shows the report "Global Phishing Survey: Trends and Domain Name Use,", Tech. Rep. 1H2014, 2014.
5. "SSAC Advisory on Fast Flux Hosting and DNS," ICANN Security and Stability Advisory Committee, Tech. Rep. SAC 025, 2008.
6. "Phishing Activity Trends Report," APWG, Tech. Rep. 2nd Quarter 2014, 2014.
7. Mohammed, reported the "Phishing: Data Mining Classification", vol. 16, no. 1, pp. 22– 29, 1990.
8. Ali Kason, published "Comparative Analysis of Machine Learning Algorithm in detection of phishing websites", vol. 19, no. 3, pp. 370–383, 2007.
9. Ricardo Pinto, implements "Artificial Neural Network for website classification with phishing characteristics", pp. 340-420, 2011
10. Wa el Hadi, reported "Detecting phishing websites using Associative classification", vol. 38, no. 11, pp. 39–41, 2014.
11. Adhi Tama, presented "A Comparative Study of Phishing Websites Classification Based on Classifier Ensembles", vol.50, pp.56-62, 2014.
12. Yan hi Du, described "Research of the Phishing Technology Based on E-mail extraction and Analysis", vol.81, pp.89-93, 2015.
13. Kumar, "Phishing Detection by Determining Reliability Factor using Rough Set Theory", vol.54, pp.57-75, 2015
14. Andric, proposed "Analysis of Phishing attacks against students", vol.39, pp.121- 125, 2016.

AUTHORS PROFILE

Dr.R.Murugeswari completed her Ph.D(CSE) from Kalasalingam University, Krishnankoil and M.E (CSE) from Arulmigu Kalasalingam College of Engineering, Krishnankoil under Anna University, Chennai. She is currently working as a Associate Professor in the Department of Computer Science and Engineering at Kalasalingam University, Krishnankoil.She has more than 18 years of Teaching Experience. Her research interests are wireless network, Evolutionary Algorithm, Big Data Analytics and Social Network Analysis. She has published more than 15 papers in various International journals and International Conferences.

G. Elizabeth Rani completed M.Tech (CSE) from Kalasalingam University and completed B.E (CSE) under Anna University. She is currently working as Assistant Professor in the Department of Computer Science and Engineering at Kalasalingam University. She has 7.5 years of Teaching Experience. Her research area interests are Image Processing, Data Mining and Data Analytics.

G.Ramasakthi pursuing her B.Tech (CSE) from Kalasalingam University, Krishnankovil.

B.Atchaiya pursuing her B.Tech (CSE) from Kalasalingam University, Krishnankovil