IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 45-48

Classification and Proactive prevention Using Fuzzy Logic and Phishing Attack Detection Data Miningalgorithm

B.Miriam Zipporahco, N.Sharmilabanu
Research Scholar Providence College for women affiliated to Bharathiar University)
Assistant Professor Providence College for women)

Abstract: *-This paper presents a design for removing phishing sites or phishingpages that are hosted probably without the knowledge of the website owner or host server. Initially the system assesses and classifies phishing emails using Fuzzy Logic and the RIPPER Data Mining algorithm. In assessing the Phishing email, Fuzzy Logic linguistic descriptors are assigned to a range of values for each key phishing characteristic indicators. The system then sends a notification to the System Administrator of the host server to indicate that it is hosting a Phishing site. The removal success rate of the identified phishing sites is 81.81% based on the notifications sent to the host of the different phishing pages.*

Keywords: *-Data Mining, Fuzzy Logic, Phishing, URL, Classification*

I. Introduction

Phishing is as an act of sending an e-mail to a user falsely claiming to be a legitimate business establishment in an attempt to scam or trick the user into surrendering private information that will be used for identity A proactive approach to minimizing phishing has been conducted where the system removes a phishing page from the host server rather than just filtering email and flagging suspected messages as spam [4].The study will take into consideration different email features in classifying phishing emails using Fuzzy Logic and Data Mining classification algorithm.

II. Related Studies

Most anti-phishing tools employ email filtering techniques to classify legitimate emails and suspected spam in the mail inbox. The user is left to decide whether to open or discard such emails. If no anti-phishing tool is installed or the user has not updated the anti-phishing program, then there is no layer of protection. This is referred to as passive anti-phishing [4]. It is because the approach only locally protects the user from a phishing attack but does not make any effort to stop or remove the Phisher at the source. The Phisher then continues with the phishing operation to further increase its victims.

While there are several email filters, browser tools, anti spyware and anti -virus software, very few research efforts have been entirely focused to protect online users from phishing attacks in the past. Existing anti-phishing and anti-spam techniques suffer from one or more limitations and they are not 100% effective at stopping all spam and phishing attacks [9]. Phishers are able to find ways to bypass existing rule-based and statistical based filters without much difficulty. Major e-mail service providers such as Yahoo, Hotmail, Gmail, and AOL filter all incoming emails separating them into Inbox (legitimate email) and junk (illegitimate email) email folders. However, these e-mail service providers do not actually attempt to remove the phishing page associated with the illegitimate email. Furthermore, Phishers have readily available tools to bypass such spam filters [5]. We refer to this as a passive anti-phishing approach.

2.1 Content Based and Non-Content Based Approach

In content based approach, phishing attacks are detected by examining site contents. Features used in this approach include keywords, spelling errors, links, password fields, embedded links, etc. along with URL and host based features[5]. Google's anti-phishing filter detects phishing and malware by examining page URL, page rank, WHOIS information and contents of a page including HTML, JavaScript, images, iframe, etc.[5]. The classifier is constantly updated to accommodate new phishing sites to cope up with the latest techniques in phishing attacks. In this approach the classifier may have higher accuracy but the result is not real -time Our approach uses Fuzzy Logic language descriptors with a range of values for each identified phishing characteristic specifically spelling errors, keywords and embedded links. The membership function for each characteristic derived as is used to assess the probability that the email is a phishing email.

Non-content based approaches are primarily based on URL and host information classification. URLs are commonly classified based on features such as URL address length and presence of special characters. Moreover, host features of URL such as IP address, site owner, DNS properties and geographical properties are

also used in the classification of Phishing emails[5]. The success rates is between 95% - 99% even in real-time processing [15].

III. Methodology

3.1 System Flow

This section describes the overall approach of the system in assessing, detecting and classifying Phishing emails. The process for notifying the hosting site or the sending site of the Phishing email is also included as well as the possible removal process. At the start, the system assesses the risk of the email using Fuzzy Logic. It then classifies the email as Phishing or legitimate email. The classification makes use of the data mining RIPPER algorithm. If the system detects that it is a phishing email, it gets the URL of the Phishing email. The host server's IP address, host server location and the contact information of the System Administrator. A notification is sent to the System Administrator of the host server informing that a phishing page is hosted by the server. The System Administrator proceeds with the removal of the Phishing page.

3.2 Detecting and Classifying Phishing Email

The proposed methodology will apply fuzzy logic and data mining algorithms to classify phishing emails based on two classification approaches such as content-based approach and non-content based approach. Specific categories or criteria are selected for each approach. The components or selected features are then identified for each category. The list of the classification approaches with the identified criteria and specific features is listed in the table below. The list will be used as basis for in the simulation and determination of phishing emails. The main characteristics of phishing emails are listed in

Table 1. Characteristics and stages of the components of phishing emails

Classification Approach	**Category/Criteria**	**Component**	**Stage/Layer**
Non-content Based Approach	URL	IP URL	Stage 1
		Redirect URL	
		Non-matching URL	
		Crawler URL	
		Long URl address	Weight = 0.5
		URL prefix/suffix	
Content-based Approach	Email Message	Spelling Errors	Stage 2
		Keywords	
		Embedded links	Weight = 0.5
Overall Weight			1.0

3.3 Use of Fuzzy Logic and RIPPER Data Mining Algorithm

The approach is to apply fuzzy logic and RIPPER data mining algorithm to assess phishing email based on the 9 identified characteristics or components. The essential advantage offered by fuzzy logic techniques is the use of linguistic variables to represent key phishing characteristic or indicators in relating phishing email probability.

3.3.1 Fuzzification and Deffuzification

During fuzzification, linguistic descriptors such as High, Low, Medium, for example, are assigned to a range of values for each key phishing characteristic indicators. Valid ranges of the inputs are considered and divided into classes, or fuzzy sets [7]. For example, redirect URL can range from 'low' to 'high' with other values in between. The degree of belongingness of the values of the variables to any selected class is called the degree of membership; Membership function is designed for each Phishing characteristic indicator. Each point in the input space is mapped to a membership value between [0, 1]. For each input the values ranges from 0 to 6 while for output, the value ranges from 0 to 100.

Deffuzification is the process of producing a measurable result in fuzzy logic given the fuzzy sets and membership degrees. It is a process in fuzzy logic where valuable data is produced from fuzzy data. This process transforms a fuzzy output of a fuzzy inference system into a crisp output [12]. Fuzzification facilitates in evaluating the rules, but the final output has to be a crisp number. The input for the defuzzification process is the collective fuzzy set and the output is a number. A useful defuzzification technique is the center of gravity. The first step of defuzzification normally removes parts of the graph to form a trapezoid. The trapezoids are then superimposed one after the other to form a single geometric shape. The centroids which is called fuzzycentroid, is calculated. The x coordinate of the centroid is the defuzzified value.

Table 2. Sample of the rule base stage 1 entries for the URL Domain and Entity Criteria

Rule #	IP URL	Redirect URL	Non-matching URL	Crawler URL	Long URL address	URL Prefix/suffix	URL Domain Entity & Criteria
1	Low	Low	Low	Low	Low	Low	Valid/Genuine
2	Low	Low	Low	Low	Low	Moderate	Valid/Genuine
3	Low	Low	Low	Moderate	Moderate	Moderate	Suspicious
4	Moderate	Low	Moderate	Low	Low	Moderate	Suspicious
5	Moderate	Moderate	Moderate	High	High	High	Fraud
6	High	High	High	High	Moderate	Moderate	Fraud

.3.2. Rule Base for Stage 2

Table 3. Sample of the rule base stage 2 entries for Email Content Domain

Rule #	Spelling Errors	Keywords	Embedded Links	Email Content Domain
1	Low	Low	Moderate	Genuine
2	Low	Moderate	Moderate	Suspicious
3	High	High	High	Fraud
4	Low	Low	Low	Genuine
5	High	Moderate	Moderate	Fraud
6	Moderate	Low	Moderate	Suspicious

IV. Results

Publicly available datasets from Phistank were used for simulation. There are two stages in determining the fuzzy data mining inference rules. 1000 sample instances are used from the Phistank archive. For rule base 1, there are 6 identified Phishing email characteristics based on the non-content based approach. The assigned weight is 0.5. For rule base 2, there are 3 identified characteristics of Phishing emails based on the content-based approach. The assigned weight is 0.5. The email rating is computed as 0.5 * URL and Domain Entity crisp (rule base 1) + 0.5 * Email Content Domain crisp (rule base 2).The RIPPER algorithm uses separate and conquer approach. It is considered an inductive rule learner that builds a set of rules that identify the classes while minimizing the amount of error. The error is determined by the number of training examples that are misclassified by the rules. The prediction accuracy is recorded in Table 4.

Table 4. Results generated from the WEKA classifier using RIPPER algorithm

applied to classify Phishing emails

Validation Mode	10 fold cross validation
Attributes	URL Domain and Entity Criteria
	Email Content Domain
Number of rules	12
Correctly classified	85.4%
Incorrectly classified	14.6%
Number of samples/instances	1000

The initial results showed that URL and Entity Domain and the Email Content Domain are important criteria for identify and detecting Phishing emails. If one of them is "Valid or Genuine", it will likely follow that the email is a legitimate email. The same is true if both of the criteria are "Valid or Genuine". Likewise, if the criteria are "Fraud", the email is considered as a Phishing email.

Table 5. Results of Phishing Pages removed after notifications were sent

Emails	**Traced Server info**	**Phishing Page Removed**	**Removal Success Rate**
23	22	18	81.81%

V. Conclusion

URL and Entity Domain as well as Email Content Domain are two important and significant Phishing criteria. If one of the criteria is "Valid or Genuine", it will likely follow that the email is a legitimate email. The same is true if both of the criteria are "Valid or Genuine". Likewise, if the criteria are "Fraud", the email is considered as a Phishing email. It should be noted, however, that even if some of the Phishing email characteristics or stage is present, it does not automatically mean that the email is a Phishing email.

References

[1]. A.-P. W. Group, "Global phishing survey: Domain name use and trends in 2h2010,"
[2]. http://www.antiphishing.org/reports/APWG, GlobalPhishingSurvey 2H2010.pdf.
[3]. A. P. W. Group, "Phishing activity trends report," 2009, http://www. antiphishing.org/reports/apwg report Q4 2009.pdf.
[4]. Anti-Phishing Working Group, "Phishing Activity Trend Report", Jan-March 2008
[5]. Shah, R.; Trevathan, J.; Read, W.; Ghodosi, H.,"A Proactive Approach to Preventing Phishing Attacks Using the Pshark Model", IEEE Sixth International Conference on Information Technology: New Generations, March 2009, pp. 915 - 921
[6]. Afroz, S.; Greenstadt, R., "PhishZoo: Detecting Phishing Websites by Looking at Them", IEEE Fifth International Conference on Semantic Computing (ICSC), 2011, pp. 368-375
[7]. Sophos, "Security Threat Report", July – 2008
[8]. Maher Aburrous, M. A. Hossain, KeshavDahal, FadiThabatah, "Modelling Intelligent Phishing Detection System for e-Banking using Fuzzy Data Mining", IEEE International Conference on CyberWorlds, 2009, pp.
[9]. 265 - 272
[10]. S. M. Bridges and R. B. Vaughn, "fuzzy data mining and genetic algorithms applied to intrusion detection," Department of Computer Science Mississippi State University, White Paper, 2001.
[11]. Rokach, Lior; OdedMaimon (2008). Data mining with decision trees: theoryand applications. World Scientific Publishing. ISBN 978-9812771711.
[12]. L. James, "Phishing Exposed," Tech Target Article sponsored by: Sunbelt software, searchexchange.com, 2006