International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-06, Issue-01, Apr 2020

Heuristic based malicious URL detection

[1]Varun Vyas, [2]Aditya Nair, [3]Allan lopes

[1,2,3]Information technology department, Universal college of engineering, Mumbai, India.

[1]mvarunphoenix999@gmail.com, [2]adityanair0202@gmail.com, [3]allan.lopes@universal.edu.in

***Abstract*— Phishing is one of the most potentially disruptive actions that can be performed on the Internet. Intellectual property and other pertinent business information could potentially be at risk if a user falls for a phishing attack . The adversary sends an email with a link to a fraudulent site to lure consumers into divulging their confidential information . One of the main goal of this research is to detect phishing attempts via email. The algorithm in the previous work analyses the body text in an email to detect whether the email message asks the user to do some action such as clicking on the link that directs the user to a fraudulent website. This work expanded the text analysis portion of that algorithm, which performed poorly in catching phishing emails. The original algorithm has considerably have a lower result in filtering out malicious email as compared to modified algorithm.To address the False Positive problem, a statistical approach was adopted and the method ameliorated the False Positive Rate while minimizing the decrease in the phishing detection accuracy.**

Keywords—phishing, email, black lists, fraud, phishing detection, malicious URL's, statistical approach.

I. INTRODUCTION

Phishing is considered as one of the malicious use of internet resources where the user are tricked into revealing their personal information, username and password and other personal information to the attacker. Phishing can appear through a variety of communication forms such as instant messaging, SMS, VOIP, online messenger, and above all the most common form of phishing attack leverages email. Fraudsters send an email to an unsuspecting user that contains a link to a domain that is seemingly legitimate in the hopes that the users will input their private information for the attacker to steal. There is no doubt phishing can be extremely damaging to all organizations since tricking a user within a business network through a phishing scam is an easy way to obtain the user's information in order to gain access to that business network. The following graph in the figure shows the number of phishing attacks per year.

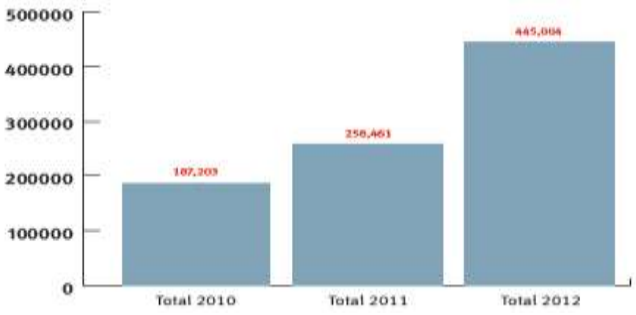

Figure 1 : Phishing Attacks per Year

The threat phishing poses to Internet users at large calls for action within the information security industry to create ways of detecting and preventing such attacks [2]. Research into the area of phishing detection has yielded several types of email analysis to determine if an email should be classified as phishing [2]. First, link or URL analysis refers to the using information about the links included within an email to detect the email used in a phishing attempt [2]. This approach helps us to check if the link in email has any kind of connection with the actual website's URL or it matches with the pattern in URL in an email in order to differentiate the features of phishing URL [2]. PhishTank, a well-known website containing a blacklist, utilizes a wisdom of crowds approach in order to collect phishing sites [2]. In order to find out whether the potential phishing sites, people report to the PishTank websites, where the phishing scams is decided by peoples vote.

II. TYPES OF ATTACKS

The attackers may direct users to specific harmful websites by misspelling URL's or by using sub-domains which take the user to a site that looks identical to the original website.

On other occasions attackers can use images in place of text to make it harder for filters to detect phishing.

A few scams include search engines where the user is directed to product sites which may offer low cost products or services. When the user tries to buy a product by entering their financial details, that information is collected by the attackers. Yet another type of attack is when the attacker is logically located in between the original website and the contaminated system. They trace

 DOI : 10.35291/2454-9150.2020.0296

details during a transaction between the legitimate website and the user.

Another technique is targeted phishing using data gathered through outside means, such as user names. The specific targets can be companies and government agencies, and the criminals send spoofed email messages misrepresenting the phishers as people from the recipient's company or organization, such as a human resources department.

Pharming is more dangerous technique in that pharmers make use of an email that simply damages the victim once the email is opened by the receiver [1]. There are many stealth application such as virus, Trojan horse, worms etc in email pharming, they get automatically downloaded or installed in their users computer, Sometimes the user may not even notice that his/her personal information are in danger until an antivirus spots the malicious applications [1]. The installed applications have a role to redirect the browser to the counterfeit sites when the user visits the official website of an organization [1]. The oblivious user provides the id and password to login the website without realizing the website is the fake webpage created by the criminal. As a result, the pharmer harvests the personal information that the victim divulges.

III. RELATED WORKS

Phishing is a criminal act which uses a combination of social engineering and technical subterfuge to steal user information [2]. The idea of "phishing" first was presented in a 1987 conference called Interex (Robson, 2011) [2]. The origin of the word "phishing" comes from the analogy that malicious Internet users lure to "fish" for credential information from the sea of Internet user by using email. In the 1996, the term "phishing" started to be used to describe the incidents that hackers were exploiting passwords from unsuspecting America On-Line (AOL) user to steal AOL accounts [2]. In today's world there are various kind of attacks which targets personal information. Stolen accounts by criminals were called "phish" by 1996, and phish started to be traded between hackers [2]. There are number of phishing attacks that are increasing exponentially and criminals are increasing the area of their activity by stealing AOL accounts to target online banking and e-commerce [2].

Blacklist Generator

This technique tries to generate an updated blacklist of phishing sites. Each web page belongs to a web site and most of them show this relation using the site's logo. Phishing pages also use legitimate site's logo to make their pages credible and claim that they belong to that site. Thus we can find which site a page claims to belong, using its logo. On the other hand, the domain of a legal site can be found by searching its name in a search engine like Google. Our technique is based on these two properties of the web pages and search engines to detect phishing pages.

Determining which pages belong to the same web site is an open problem, although some heuristic approaches have been proposed . With the same host name we can approximate each website by all the pages. It is a useful technique but not quite accurate. The output of Blacklist Generator is an XML document.. This experiments indicates that in 74.4% of cases, searching a company name in Google brings its web site's link as the first search result item and it is beneficial to suggest this item as alternatives for users when their access to a phishing site is blocked.

In most cases, the starting point of a phishing attack is an email. Users receive emails that contain suspicious links that direct them to phishing sites. The link to all phishing sites can be found in the body of the email, so emails are a valuable source to make a blacklist of phishing sites. In addition, since our algorithm is time-consuming, it can be useful to Internet users who can tolerate due delays for their increased safety. So the best place to apply our algorithm is to an email server. Figure 2 shows proposed architecture for the blacklist generator[8].

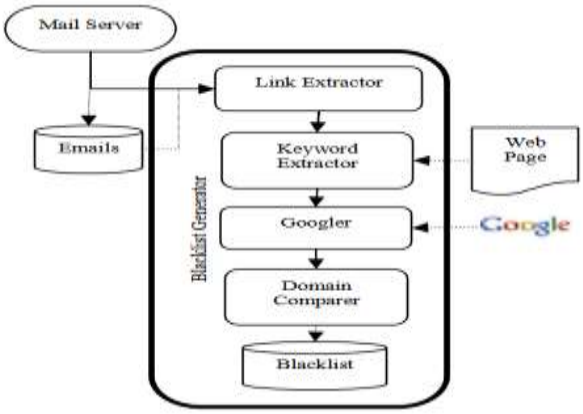

Figure 2: Architecture of blacklist generator

IV. ANALYSIS RESULTS

Constructed the hierarchical structure graphs using around 4,000 data lists of the malicious Web sites in the way mentioned in the previous section. Then, we counted up values of the degree with respect to AS numbers, IP address blocks, IP addresses, and registrars. The proportions of the ingredients with the degree values that are more than10 are around 3%, 1.5%, 3.5%, and 20% respectively. We assume that the data with high degree values are heavily related to the deployments of malicious Web sites. Next, we show the averages and variances of the degree values of each kind of the domain information[9]. While the AS, IP address block, and IP address elements do not have so high average values of the degree, the registrar element has a high average value of it[9]. This would be due to that there are not so much registrars in the wild, so the number of the use of each

 DOI : 10.35291/2454-9150.2020.0296

registrar could be large. Likewise, from the results of the variances, we can see that the variance of the degree value of the registrar is very large. This is because attackers intentionally and intensively use specific registrars. In fact, some researches including our previous research reported that attackers use specific registrars intensively . The variance of the IP address is also rather large, and this is due to that some of attackers tend to change domain names in a short period for avoiding blacklists and which results in assigning many domains to one IP address. Similarly, the variance of the AS is rather large, and this would be because the number of IP address blocks that each AS owns is different from each other[9]. On the other hand, the IP address block has relatively low value of variance[9].

Disadvantages of Black list approach

Blacklist only accounts known variable so it can protect it from identifies threats.

Malware often exploits blacklists that are designed to evade detection specifically.

Blacklists can be best if you aren't concerned with protecting a certain system. The system mainly contains public, non-sensitive, information..

There are instances when this type of service results in false positives. The updating process of the list is most probably slow, therefore if there is a new entry of phishing website may prove harm because it has not been added to blacklist.

V. DEVELOPED APPLICATION

The heuristic-based phishing sites detection technique analyzes and extracts phishing site features from the URL such as Domain, PrimaryDomain, SubDomain, PathDomainand detects phishing sites using the information obtained from the features. With heuristic approach, a signature database of known attacks is built to scan a web page. The websites will be considered as phishing websites if the heuristic patterns of the websites match signatures in the database. This approach can detect new phishing sites and temporary phishing sites because it extracts features from the requested web page.

There is an interface which is provided by the system where the user can write his/her query. Once the search button is clicked, it gives the list of URLs on the same page. A URL is a protocol that is used to indicate the location of data on a network. The URL is composed of the protocol, subdomain, primary domain, top-level domain (TLD), and path domain. In meantime; it saves all the URLs in the database. The protocol refers to a communication protocol for exchanging information between information devices; e.g., HTTP, FTP, HTTPS, etc. Protocols are of various types and are used in accordance with the desired communication method. All the total score in the database are calculated and all are mentioned in eighteen factors for each URL. Then based on the total_score value of each URL, they are rearranged in the descending order, which means if URL has high total_score value then it will appear as the top most result and accordingly rest comes as per their total_score value.

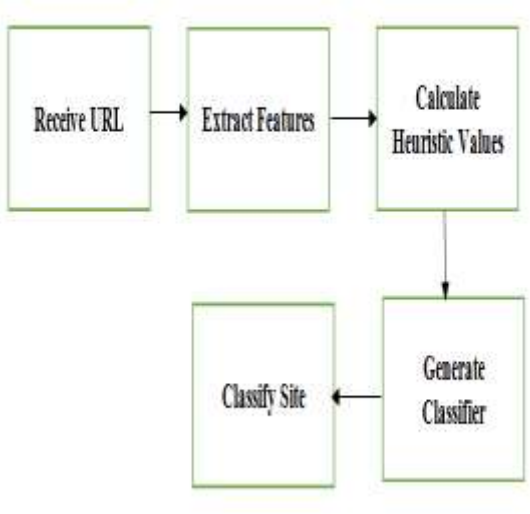

Figure 3 : Calculation based on heuristic approach

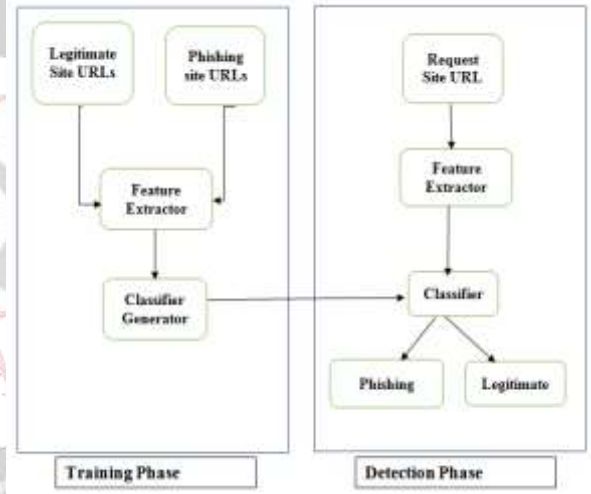

Figure 4 : The data undergoing the training phase and detection phase

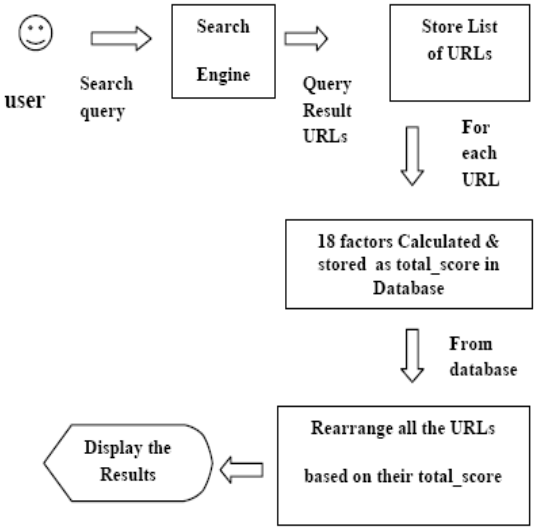

Figure 5: Procedures in getting the result

 DOI : 10.35291/2454-9150.2020.0296

VI. EVALUATION

We have selected 9,661 phishing sites from PhishTank as training dataset and testing dataset that contains 1,000legitimate sites from DMOZ and 1,000 phishing sites from PhishTank. Experimental procedure is divided into 5 phases.

Phase 1

Dataset is collected from PhishTank[7] and imported into MYSQL with 9,661 phishing websites.

B. Phase 2

Four features (Primary Domain, Sub Domain, Path Domain and Domain) are selected. In this phase, we use PHP to select four features as Primary Domain, Sub Domain, Path Domain and Domain from URLs in dataset.

phish_id	domain	primarydomain	subdomain	pathdomain
1777670	bishopfitaherlyassembly.org	bishopfitaherlyassembly		paypal.paypal.priv8.us.2012.pp.login.h
1777669	usenmedente.com.tr	usenmedente		media.patrick.DanMaker
1777668	ausadia.com	ausadia	paypal	
1777666	ausadia.com	ausadia	paypal	
1777662	dawhoo.com	dawhoo		wp.content.doc9le

Table1

C. Phase 3

In this phase, search engine spelling suggestions andalexa.com are used to calculate the value of the heuristics.

WEIGHT OF THE HEURISTICS

Heuristic	Phishing sites	Weight
PrimaryDomain	2971	0.105
SubDomain	1380	0.049
PathDomain	3787	0.134
PageRank	7514	0.266
AlexaRank	6437	0.227
AlexaReputation	6208	0.219

Table 2

D. Phase 4

In this phase, we calculate the value of "vs" for each URL from dataset of 9,661 phishing sites and compare to the thresholds.

RESULT OF CLASSIFYING WEBSITES

Threshold	Phishing sites	Ratio
−0.2	4830	50%
−0.1	5120	53%
0	6122	63%
0.1	6768	70%
0.2	7539	78%
0.3	8696	90%
0.4	9182	95%
0.5	9374	97%
0.6	9661	100%

Table 3

E. Phase 5

In this phase, our proposed technique is tested with testing dataset which contains 1,000 phishing sites from PhishTank[7]. and 1,000 legitimate sites from DMOZ. In case of 2-classprediction, there are four possible results that are defined as follows:

•True Positive (TP): If the result of prediction is legitimate site and the actual value is also legitimate site.

•False Positive (FP): If the result of prediction is phishing site but the actual value is legitimate site.

•True Negative (TN): If the result of prediction is phishing site and the actual value is also phishing site.

•False Negative (FN): If the result of prediction is legitimate site but the actual value is phishing site[10].

If the accuracy ratio is calculated as follows: Accuracy ratio = (TP+TN)/(TP+TN+FP+FN), the results of the test will be shown in Table IV. From the obtained results, we have found that this technique has a high accuracy rate of 97% with the threshold value of 0.5[7][10].

Threshold	TP	TN	FP	FN	Accuracy Ratio
−0.2	623	497	377	503	56%
−0.1	576	602	424	398	59%
0	611	627	389	373	62%
0.1	587	634	413	366	61%
0.2	675	705	325	295	69%
0.3	795	868	205	132	83%
0.4	889	945	111	55	92%
0.5	979	967	21	33	97%
0.6	876	997	124	3	94%

Table 4: Result of testing

VII. FUTURE SCOPE

Attempts will be made to change the application installation such that it can be received by potential clients remotely.

VIII. CONCLUSION

In this paper we proposed a heuristic-based phishing detection technique that employs URL-based features. The method combines URL-based features used in previous studies with new features by analysing phishing site URLs. Additionally, we generated classifiers through several machine learning algorithms and determined that the best classifier was random forest. It showed a high accuracy of 98.23% and a low false-positive rate. The proposed technique can provide security for personal information and reduce damage caused by phishing attacks because it can detect new and temporary phishing sites that evade existing phishing detection techniques, such as the blacklist-based technique. In future work, we intend to address the time-intensive disadvantage of the heuristic-based technique. With a large number of features, it is time-consuming for the heuristic based approach to generate classifiers and perform classification. Therefore, we will apply algorithms to reduce the number of features

 DOI : 10.35291/2454-9150.2020.0296

and thereby improve performance. In addition, we will examine a new phishing detection technique that uses not only URL-based features.

REFERENCE

[1] URL Based Phishing Detection System using Machine Learning by Prof. Prashant Rathod, Usama Khatri, Kinjol Shah.

[2] Text-Based Phishing Detection Using a Simulation Model by Gilchan Park.

[3] T. Vyas, P. Prajapati and S. Gadhwal, "A survey and evaluation of supervised machine learning techniques for spam e-mail filtering," 2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, pp. 1-7, 2015.

[4] Arati M. Dixit Anjali B. Sayamber. Malicious url detectionand identification.International Journal of Computer Appli-cations (0975 8887) Volume 99 No.17, August 2014, 2014.

[5] Adrian-Stefan Popescu, Dumitru-Bogdan Prelipcean, and Dragos Teodor Gavrilut. A study on techniques for proac-tively identifying malicious urls. In17th International Sym-posium on Symbolic and Numeric Algorithms for ScientificComputing, SYNASC 2015, Timisoara, Romania, September21-24, 2015, pages 204–211, 2015.

[6] Yue Zhang, Jason I. Hong, and Lorrie Faith Cranor. Cantina:a content-based approach to detecting phishing web sites. InProceedings of the 16th International Conference on WorldWide Web, WWW 2007, Banff, Alberta, Canada, May 8-12,2007, pages 639–648, 2007.

[7] Phishtank :- www.phishtank.com

[8] Mohsen Sharifi , Seyed Hossein Siadati "A phishing sites blacklist generator" April 22 2008.

[9] Yoshiro Fukushima, Yoshiaki Hori ,Kouichi Sakurai "Proactive Blacklisting for Malicious Web Sites by Reputation Evaluation Based on Domain and IP Address Registration" 03 January 2012.

[10] Luong Anh Tuan Nguyen, Ba Lam To, Huu Khuong Nguyen, Minh Hoang Nguyen "Detecting phishing web sites: A heuristic URL-based approach" 06 January 2014.

 DOI : 10.35291/2454-9150.2020.0296