“Phish Mail Guard :Phishing Mail Detection Technique by using Textual and URL Analysis”

Jayshree Hajgude
Department of Information Technology
V.E.S.Institute of Technology.
Mumbai, India.
jayshreehajgude@gmail.com

Lata Ragha
Department of Computer Engineering
Terna Engineering College.
Mumbai, India.
latha.ragha@gmail.com

Abstract — **Phishing is the combination of social engineering and technical exploits designed to convince a victim to provide personal information, usually for the monetary gain of the attacker. Phishing emails contains messages to lure victims into performing certain actions, such as clicking on a URL where a phishing website is hosted, or executing a malware code. Phishing has become the most popular practice among the criminals of the Web. Phishing attacks are becoming more frequent and sophisticated. URL and textual content analysis of email will results in a highly accurate anti phishing email classifier. We propose a technique where we consider the advantages of blacklist, white list and heuristic technique for increasing accuracy and reducing false positive rate. In heuristic technique we are using textual analysis and URL analysis of e-mail. Since most of the phishing mails have similar contents, our proposed method will increase the performance by analysing textual contents of mail and lexical URL analysis. It will detect phishing mail if DNS in actual link is present in blacklist.DNS is present in white list then it is considered as legitimate DNS. If it is not present in blacklist as well as white list then it is analyzed by using pattern matching with existing phishing DNS, contents found in mail and analysis of actual URL. With the help blacklist and white list we are avoiding detection time for phishing and legitimate email. At the same time we are decreasing false positive rate by combining features of DNS, textual content analysis of email and URL analysis.**

I. INTRODUCTION

Phishing is the criminally fraudulent process of attempting to acquire sensitive information such as usernames, passwords and credit card details by masquerading as a legitimate trusted by customers in an electronic communication. Communications purporting to be from banks, online organizations, internet services providers, online retailers, insurance agencies and so on. Popular social web sites (YouTube, Face book, MySpace, Windows Live Messenger), auction sites (eBay), online banks (Wells Fargo, Bank of America, Chase), online payment processors (PayPal), or IT Administrators (Yahoo, ISPs, corporate) are commonly used to lure the users. Phishing is typically carried out by email, and it often directs users to enter details at a fake website which is almost identical to the legitimate one. Even using server authentication, it still requires skill to detect that the website is malicious. Phishing is an example of social engineering techniques used to deceive users, and exploits the poor usability of current web security technologies. Attempts to deal with the growing number of reported phishing incidents include legislation, user training, public awareness, and technical security measures. Phishing is the process of fooling a consumer into divulging personal information, such as credit card numbers or passwords, usually by sending an email carefully constructed to appear as if it's from a bank or other trusted entity, such as PayPal. As people increasingly rely on the Internet for business, personal finance and investment, Internet fraud becomes a greater and greater threat. Internet fraud takes many forms, from phony items offered for sale on eBay, to scurrilous rumors that manipulate stock prices, to scams that promise great riches if the victim will help a foreign financial transaction through his own bank account. One interesting species of Internet fraud is phishing. Phishing attacks use email messages and web sites designed to look as if they come from a known and legitimate organization, in order to deceive users into disclosing personal, financial, or computer account information. Phishing emails usually contain a message from a credible looking source requesting a user to click a link to a website.

However, phishing has become more and more complicated and sophisticated so that phishers can bypass the filter set by current anti-phishing techniques and cast their bait to customers and organizations. A possible solution is to create a robust classifier to enhance the phishing email detection and protect customers from getting such emails. By analyzing phishing emails, it is observed that phishing emails often include certain phrases, for example, “security”, “verify your account”, “if you don’t update your details within 2 days, your account will be closed”, “click here to access to your account” and so on. These phrases may appear in the “subject:” line in an email or email content. Therefore, most phishing emails are largely similar in wording, especially the most important terms, such as “security”, “expire”, “unauthorized”, “account”, “login”, etc. Such terms are useful to classify if an email is a phishing email [1][2]. In addition, Phishing emails often alert customer to click links to other websites which the real link is not the same as it is shown in the pages. In the proposed method we are using hybrid method for phishing mail detection which is a combination of blacklist, white list and heuristic technique. In heuristic technique we are considering textual and URL analysis for further classification. Hybrid email classification is used to enhance the classification accuracy of email messages. A number of

978-1-4673-4805-8/12/$31.00 ©2012 IEEE

features are extracted from email messages like text content, DNS name from visible link, URL features. This results into representing each message as a set of values where each value shows existence of that feature in that e-mail.

In this paper section II describes background and related work on Phishing mail detection methods and their drawbacks. Section III includes phishing mail detection workflow. Section IV deals with proposed phishing mail detection technique. Section V contains conclusion.

II. BACKGROUND AND RELATED WORK

In this paper, we assume that phishers use e-mail as their major method to carry out phishing attacks **.**

A. Phishing Attack life cycle

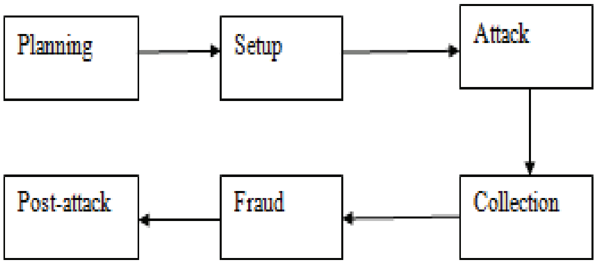

Figure 1.Phishing Attack LifeCycle

Phishing attack lifecycle is depicted in figure 1.Phishing attack lifecycle can be decomposed in Planning, Setup, Attack, Collection, Fraud and Post-Attack Actions[4].

Planning: Phishers decide which business to target and determine how to get e-mail addresses for the customers of that business. They often use the same mass-mailing and address collection techniques as spammers.

Setup: Once they know which business to spoof and who their victims are, phishers create methods for delivering the message and collecting the data. Most often, this involves e-mail addresses and a Web page.

Attack: This is the step people are most familiar with the phisher sends a phony message that appears to be from a reputable source.

Collection: Phishers record the information victims enter into Web pages or popup windows.

Identity Theft and **Fraud:** The phishers use the information they've gathered to make illegal purchases or otherwise commit fraud. As many as a fourth of the victims never fully recover.

Phishers try to send an e-mail to a user falsely claiming to be an established legitimate enterprise in an attempt to scam the user into surrendering private information that will be used for identity theft. The e-mail directs the user to visit a Web site where they are asked to update personal information, such as passwords and credit card, social security, bank account numbers that the legitimate organization already has.

B. Types of phishing attack techniques

Figure 2 shows different phishing attack techniques

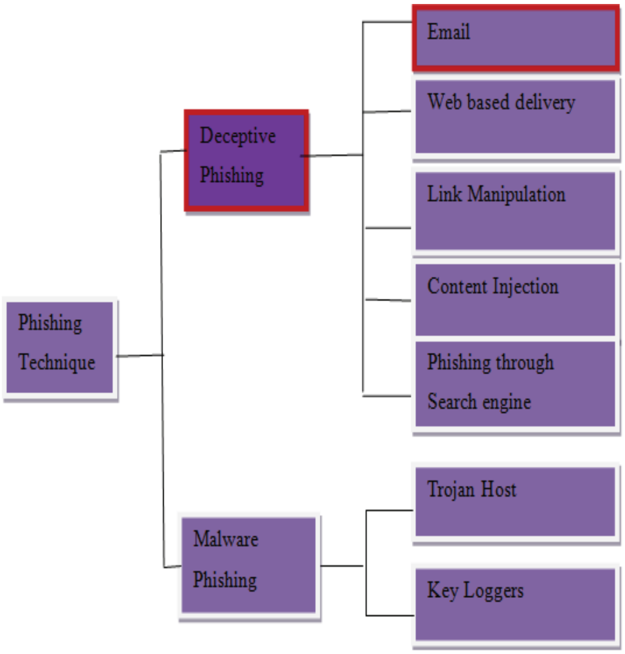

Figure 2. Different types of Phishing attack.

Two different types of phishing attacks may be distinguished [3].

1. Malware-based phishing

For malware-based phishing is a malicious software is spread by deceptive emails or by exploiting security holes of the computer software and installed on the user's machine. Then the malware may capture user input, and confidential information may be sent to the phisher.

Trojan hosts are invisible hackers trying to log into your user account to collect credentials through the local machine. The acquired information is then transmitted to phishers.

Key loggers refer to the malware used to identify inputs from the keyboard. The information is sent to the hackers who will decipher passwords and other types of information. To prevent key loggers from accessing personal information, secure websites provide options to use mouse click to make entries through the virtual keyboard.

2. Deceptive phishing.

Deceptive phishing, in which a phisher sends out deceptive emails pretending to come from a reputable institution, e.g. a bank. In general, the phisher urges the user to click a link to a fraudulent site where the user is asked to reveal private information, e.g. passwords. This information is exploited by the phisher, e.g. by withdrawing money from the users bank account. There are different techniques used in deceptive phishing, Phishers may send the same email to millions of users, requesting them to fill in personal details.

These details will be used by the phishers for their illegal activities. Phishing with email and spam is a very common phishing scam. Most of the messages have an urgent note which requires the user to enter credentials to update account information, change details, and verify accounts. Sometimes, they may be asked to fill out a form to access a new service through a link which is provided in the email.

Web based delivery is one of the most sophisticated phishing techniques. Also known as "man-in-the-middle," the hacker is located in between the original website and the phishing system. The phisher traces details during a transaction between the legitimate website and the user. As the user continues to pass information, it is gathered by the phishers, without the user knowing about it.

Content injection is the technique where the phisher changes a part of the content on the page of a reliable website. This is done to mislead the user to go to a page outside the legitimate website where the user is asked to enter personal information.

Phishing through Search Engines Some phishing scams involve search engines where the user is directed to products sites which may offer low cost products or services. When the user tries to buy the product by entering the credit card details, it's collected by the phishing site. There are many fake bank websites offering credit cards or loans to users at a low rate but they are actually phishing sites.

Link manipulation is the technique in which the phisher sends a link to a website. When the user clicks on the deceptive link, it opens up the phishers website instead of the website mentioned in the link. One of the anti-phishing techniques used to prevent link manipulation is to move the mouse over the link to view the actual address.

C. Phishing mail detection Techniques:

Figure 3 shows different types of phishing detection techniques. Blacklists and heuristic are arguably the most popular phishing detection techniques. As evaluated in, although blacklists achieve low false positives, their detection rates suffer at zero-hours and are evaluated to detect only 20% of zero-hour phishing attacks. Heuristics of mail are able to constantly detect phishing attacks at a similar rate. However, heuristics were evaluated to have high false positives.

Effectiveness of a blacklist-based solution depends on the time it takes until a phishing site is included. This is because many phishing pages are short-lived and most of the damage is done in the time span. The techniques are described in detail below .The suspicious URL is matched against a list of known Phishing sites. This method is susceptible to "zero day attacks". Also, techniques like URL obfuscation and routing through alternate domain name can hinder this method ineffective. In this form of attack the URL's host contains a valid looking domain name, and the path contains valid looking domain name, and the path contains the organization being phished.

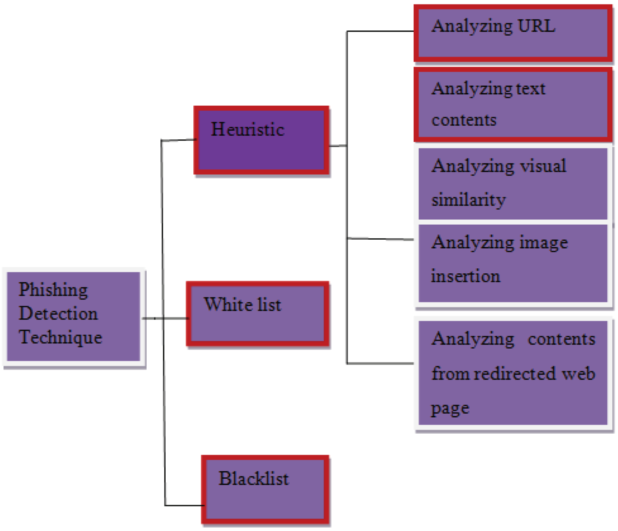
Phishing Detection Technique
Heuristic
White list
Blacklist
Analyzing URL
Analyzing text contents
Analyzing visual similarity
Analyzing image insertion
Analyzing contents from redirected web page

Figure 3. Phishing Mailing Detection Techniques

Most of the heuristics used are subjective and produce a large number of false positives. This solution is not limited to URL processing, but also analyzes the page layout. Although some heuristics are used in this solution, they are used only in the pre-processing stages, and the actual phish detection is completely independent of them. Drawbacks of this method are as mentioned below:

D. Literature Survey

There were lots of methods proposed for phishing mail detection .However, false positives have been observed in these methods. Also, a web site routed through content distribution network would create problems for domain based checks. As evaluated in [5][6], although blacklists achieve low false positives, their detection rates suffer at zero-hours and are evaluated to. detect only 20% of zero-hour phishing attacks. On the other hand, heuristics are able to constantly detect phishing attacks at a similar rate. However, heuristics were evaluated to have high false positives.

Phish Block is a hybrid phishing technique which is a combination of blacklist and heuristic approach and is explained in [7]. Lookup based systems suffer from high false negatives while classifier systems suffer from high false positives. To better detect fraudulent websites, we propose in this work an efficient hybrid system that is based on both lookup and a support vector machine classifier that checks features derived from websites URL, text and linkage this method is very complex and tested for small dataset.

Next technique introduced was Phish Catch. Phish Catch is a heuristic based algorithm which will detect phishing emails and alert the users about the phishing emails [8]. The phishing filters and rules in the algorithm are formulated after extensive research of phishing methodologies and tactics. Phish catch rate of this technique is less.

John Yearwood, Musa Mammadov and Arunava Banerjee have proposed a heuristic method called profiling phishing email based on hyperlink information. This technique uses hyperlinks in the phishing emails as features and structural properties of emails along with whois information on hyperlinks as profile classes. But drawbacks of this method are for blacklisted url it is time consuming. No valid criterion for measuring the importance of the classes present in profiling [9].

PILLER is a machine learning based approach to e-mail classification [10]. The tool decides that whether some communication is deceptive, that is whether it is designed to trick the user into believing they are communicating with a trusted source, when in reality the communication is from an attacker. The decision is based on information from within the email or feature vector itself combined with information from external sources.

Bergholz, De Beer, Glahn, Moens, Gerhard and Strobel proposed a Machine Learning classifier with model-based features that is, features that themselves are classification models and require to be trained first prior to their use by a parent classifier [11]. The proposed classifier used a total of 27 features, two of which were model-based features.

Jeong-Ho Chang proposed a technique called Improved Phishing Detection using Model-Based Features [12]. This technique uses heuristic technique for phishing mail detection. But drawbacks of this method are low accuracy and blacklist not considered.

Chandrasekaran proposed a technique to classify phishing based on structural properties of phishing emails [13].They have used a total of 25 features mixed between style markers (e.g. the words suspended, account, and security) and structural attributes, such as the structure of the subject line of the email and the structure of the greeting in the body.

Lexical URL Analysis for Discriminating Phishing and Legitimate E-Mail Messages is proposed in [14]. In proposed method we are try to minimize false positive rate. Andrew Jones proposed Lexical URL Analysis for Discriminating Phishing and Legitimate E-Mail messages. The centre claim of this paper is that lexical URL analysis technique can enhance the classification accuracy of email classifiers.

Liping Ma, Bahadorrezda Ofoghi , Paul Watters, Simon Brown proposed a method Detecting Phishing Emails Using Hybrid Features[15].Author presented a robust classifier to detect phishing emails using hybrid features and to select features using information gain. They have used 10 cross-validations to build an initial classifier which performs well.

Ripan Shah, Jarrod Trevathan, Wayne Read and Hossein Ghodosi proposed a Proactive Approach to Preventing Phishing Attacks Using a Pshark[16] .This paper proposes a proactive approach to remove a phishing page from the host server.

In this paper we have proposed a method which is a combination of blacklist, white list and heuristic to detect phishing mail.

III Phishing Mail Detection

*A.*Worlflow of Phishing Mail Detection

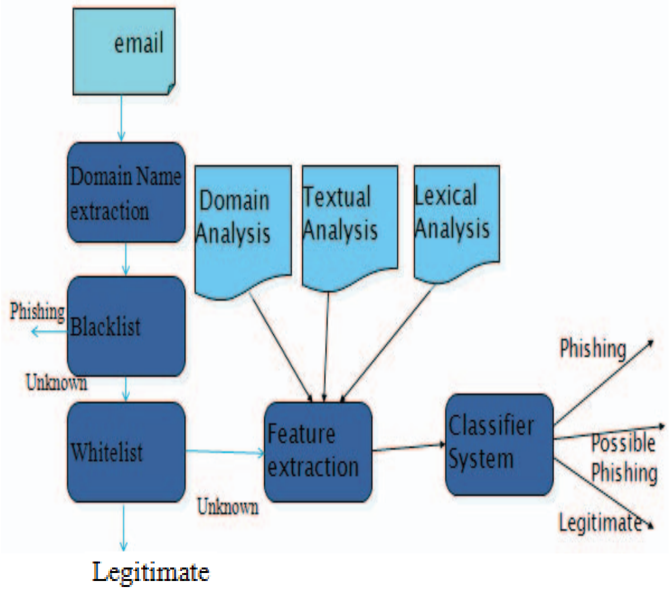

Figure 4. Workflow of Phishing Mail Detection

Workflow of proposed method is shown in figure 4.In the proposed method we are using hybrid method for phishing mail detection which is a combination of blacklist, white list and heuristic technique. In heuristic technique we are considering textual and lexical URL analysis for further classification. Hybrid email classification is used to enhance the classification accuracy of email messages. A number of features are extracted from email messages like text content, DNS name from visible link, URL features.

B. Modules used in Phishing Mail Detection

Proposed method mainly includes three Modules DNS analyzer, Classifier system, Lookup System. DNS analyzer component checks e-mail is phishing or not phishing by analyzing visual DNS and actual DNS in e-mail. This module checks the DNS of hyperlink is in Black list and White list respectively. If it is present in blacklist then phishing mail warning will be given to the user. If it is present in white list then it is considered as legitimate mail. If it is not present in blacklist or white list then it calls pattern matching module. This module is implemented using AnalyzeDNS algorithm.

As a part of lookup system we are maintaining blacklist and white list. Black list stores list of known fake DNS, while the white list contains list known valid or registered DNS.DNS analyzer uses these list for checking whether the domain in visual link is present in black list or white list.

Lookup systems typically have high precision since they are less likely to consider authentic sites as fake. They are also easier to implement than classifier systems. However, lookup systems are more susceptible to higher levels of false negatives .we are maintaining list of blacklisted domains as well as legitimate domain.

Classifier system is used to analyze mail based on heuristic features like URL features from the link, email body features, content features, email header features etc. In proposed classifier system mainly we are using URL features and textual features for performing heuristic analysis of mail.

AS given in reference [17][18][19] URL analysis plays very important role in phishing mail detection. We are using 7 features from URL feature and LUA value for lexical URL analysis. As per study in [20] empirical evaluation for feature selection following feature has maximum weight. So we have selected these 6 features for our proposed method.

IV. Phishing Mail Detection Algorithm

Phishing mail detection works by analyzing the blacklist and white list checking, difference between the visual link and the actual link, textual analysis and lexical analysis. There are different techniques to detect phishing emails that uses email hyperlink properties, e-mail header analysis, file attachment scanning etc. We developed a algorithm to detect phishing emails. It also calculates the similarity of a URL with a trusted site, performs textual analysis and lexical URL analysis.

```
R1=analyzetext(emailtext)
For every link in the email following Algorithm Executes:
v_link : visual link            a_link :  actual link
PhishmailDetect (v_link , a_link)
{
v_dns=GetDNS(v_link)       //DNS name from visual  link
a_dns=GetDNS(a_link)       // DNS name from actual  link
R2=analyzeDNS(actuallink)
R3=analyzelexicalurl(alink)
if (a_dns exists in blacklist)  return phishing
else if ((a_dns exists in whitelist) or( R1==false and
R2==false and R3==false)  return not phishing
else if (R1==true and  R3==true) then return phishing
else
 if (R1 or R2 or R3 is true) then  Possible phishing
}
```

AnalyzeDNS Algorithm:

Analyze the actual DNS as whether it is blacklisted or white listed or if unknown. Depending on the result it gives output. If it is not present in blacklist or white list then it will check for pattern matching.

Pattern matching algorithm compares sender dns and actual dns. If actual DNS and Sender DNS is different then return phishing. If it is same then actual DNS is compared with each element in SEED_SET database. Depending on the result gives user warning message.

Analyze_text algorithm analyzes text and keeps track of number of blacklisted tokens from the mail contents if number of blacklisted tokens present are more than threshold then it is considered phishing and returns true .

Similarly we analyze URL with respect to different features and try to analyze links embedded in the email are phishing links or legitimate links.

```
boolean Analyzetext(emailtext)
{
 Set tcount=0
 For(every token in email)
 If (token belongs to blacklisted tokens)
  tcount=tcount+1
  If(tcount >threshold)
           Return  true
   Else
            Return false
 }
```

From literature survey we have understood that following are the features which plays very important role in identifying phishing URL. Analyze URL algorithm converts link URL into tokens and keeps count of presence of black listed features from URL. If count is greater than threshold then it returns true else it return false.

```
boolean AnalyzeURL(url)
{
   Convert url into set of tokens
   if (noof@symbol>0)
   count=count+1
   if (noofwordsllinks>0)
   count=count+1
   if (noofperiods>2)
   count=count+1
   if (noofdomains>2)
   count=count+1
   if(length(hostname)>22)
   count=count+1
   if(number_ of_ dash(hostname)>2)
   if (URL contains IP address)
   count=count+1
  if (count> threshold)
            Return true
   else
            Return false
   }
```

V. CONCLUSION

We have studied different fishing attacks on email. We have described different phishing mail detection technique. A hybrid method has been proposed to detect phishing mail which is a combination of blacklist, white list and heuristic method. In heuristic detection technique we are considering textual analysis of email and lexical analysis of email for detection. This mechanism can effectively detect phishing mails as compared to the previous methods. This mechanism uses combination of textual analysis and lexical URL analysis. From previous study it is understood that most of the phishing mails has similar text. So with the help of textual analysis we can effectively determine phishing mail. For increasing effectiveness of mechanism we are using lexical URL analysis. Our main aim is to reduce false positive rate. So by analyzing DNS from the link, textual contents of mail and URL analysis we are trying to reduce false positive rate. At the same time we are taking care of possibility of phishing email by alerting user with possible phishing message.

References

[1] Danesh Irani, Steve Webb, Jonathon Giffin and Calton Pu, “Evolutionary Study of Phishing”, IEEE International Conference on Web Security, pp. 206-210, 2008.

[2] Cynthia Dhinakaran and Jae Kwang lee, “Reminder: please update your details”, Phishing Trends IEEE first International Conference on Networks & Communications, pp. 295-300, 2009.

[3] Jasveer Singh, “Detection of phishing emails”, International Journal of Computer Science and Technology - IJCST, Vol.2, Issue 3, pp. 547-549, September 2011

[4] Huajun Huang, Shaohong Zhong, Junshan Tan, “Browser-side Countermeasures for Deceptive Phishing Attack” ,IEEE fifth International Conference on Information Assurance and Security, pp.352-355,2009.

[5] A. Alnajim and M. Munro, “An evaluation of user’s anti-phishing knowledge retention”, IEEE International conference on Information Management and ICIME ’09, pp. 210-214,April 2009.

[6] S.Sheng, B.Wardman, G.Warner, L.F.Cranor, J.Hongand C. Zhang.. “An empirical analysis of phishing blacklists”, Sixth International Conference on Email and AntiSpam, July 16-17, 2009.

[7] Hossom ,M. A. Fahmy and salma A. Ghoneim , “PhishBlock: A hybrid anti-phishing tool”, International Conference on Communications, Computing and Control Applications, IEEE Digital Library, pp. 1-5, March 2011.

[8] Nargundkar S. and Trithani N. ,”Phishcatch: Phishing Detection tool” , 33rd IEEE International Conference on Computer Software and Application, pp.451-456, 2009.

[9] John Yearwood, Musa Mammadov and Arunaya Bannerjee ,"“Profiling Phishing Emails Based on Hyperlink Information”, International Conference on Advances in Social Networks Analysis and Mining, pp. 1-10, 2010.

[10] Fettee N. Sadeh and A. Tomasic ,” Learning to detect Phishing email”, Proceedings of the 16th international conference on World Wide Web, Published in ACM digital library, pp.649–656, New York, USA 2007.

[11] A. Bergholz , J. De Beer , S. Glahn , M.F. Moens, Gerhard , P. P.,and S. Strobel, “New filtering approaches for phishing email”, Journal of Computer Security, vol. 18, pp. 7–35, January 2010.

[12] Jeong-Ho Chang, “Improved Phishing Detection using Model-Based Features”, IEEE First International Conference on Networks & Communications, pp 295-300, 2009.

[13] Chandrasekaran, M., Narayanan, K., and Upadhyaya, S., “Phishing E-mail Detection Based on Structural Properties”, New York State Cybersecurity Conference Symposium on Information Assurance: Intrusion Detection and Prevention, pp. 2-8. 2006.

[14] Mahmoud Khonji and Youssef Iraqi, “Lexical URL analysis for discriminating phishing and legitimate email”, 6th IEEE International Conference on Internet Technology and Secure Transaction, pp.422-427, 2011.

[15] Liping Ma, Bahadorrezda Ofoghi, Paul Watters, Simon Brown, "Detecting Phishing Emails Using Hybrid Features," uic-atc, pp.493-497, Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing, 2009 .

[16] Ripan Shah, Jarrod Trevathan, Wayne Read and Hossein Ghodosi, “Proactive Approach to Preventing Phishing Attacks Using a Pshark”, Sixth International Conference on Information Technology: New Generations, pp. 915-921, 2009.

[17] Ma, J., Saul, L., Savage, S., and Voelker, G., “Identifying Suspicious URLs: An Application of Large-Scale Online Learning”, Proceedings of the 26th International Conference on Machine Learning, Montreal, Canada, 2009

[18] Brad Wardman, Gaurang Shukl and Gary Warner, ”Identifying Vulnerable Websites by Analysis of Common Strings in Phishing URLs” , IEEE Conference eCrime ResearchersSumit,pp.1-13,2009.

[19] Anh Le, Athina Markopoulou ,Michalis Faloutsos,”PhishDef URL Names Say It All, IEEE Conference INFOCOM , pp. 191-1 95 , 2011.

[20] Fergus Toolan and Joe Carthy, “Feature Selection for Spam and Phishing Detection”,□ □ □ IEEE International Conference eCrime Researchers Summit, pp.1-12, 2010