Detection of Phishing Websites Using Data Mining Techniques

Anindita Khade
Dept. of CE ,
Lokmanya Tilak College of Engineering
Koparkhairane,Navi Mumbai 421302

Dr. Subhash K Shinde
Dept of CE,
Lokmanya Tilak College of Engineering
Koparkhairane,Navi Mumbai 421302

Abstract—Detecting any Phishing website is really a complex and dynamic problem involving many factors and criteria. Because of the ambiguities involved in phishing detection, fuzzy data mining techniques can be an effective tool in detecting phishy websites.In this paper we propose a method which combines fuzzy logic along with data mining algorithms for detecting phishy websites. Here, we define 3 different phishing types and 6 different criteria for detecting phishy websites with a layer structure. We have used RIPPER data mining algorithm for classification. Furthermore, after the email has been assessed and classified as a Phishing email, the system proactively gets rid of the Phishing site or Phishing page by sending a notification to the System Administrator of the host server that it is hosting a Phishing site which may result in the removal of the site. Furthermore, after classifying the Phishing email, the system retrieves the location, IP address and contact information of the host server.

***Keywords*: Phishing, Ripper algorithm, fuzzy logic**

1. INTRODUCTION TO PHISHING

Phishing websites are forged websites that are created by malicious people to appear as a real websites. 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 theft. The impact is the breach of information security through the compromise of confidential data and the victims may finally suffer losses of money or other kinds. There were at least 67, 677 phishing attacks reported by the Anti-Phishing Working Group (APWG) in the last six months of 2010. The latest reports showed that most phishing attacks are “spear phishing” that aim the financial, business and payment sectors. E-banking Phishing website is a very complex issue to understand and to analyze, since it is joining technical and social problem with each other for which there is no known single silver bullet to entirely solve it. The motivation behind this study is to create a resilient and effective method that uses Fuzzy Data Mining algorithms and tools to detect phishing websites in an automated manner. 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. DM approaches such as neural networks, rule induction, and decision trees can be a useful addition to the fuzzy logic model.

2. RELATED WORK

Intrusion detection is software, hardware or combination of 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. Phishing website is a recent problem, nevertheless due to its huge impact on the financial and on-line retailing sectors and since preventing such attacks is an important step towards defending against e-banking phishing website attacks, there are several promising approaches to this problem and a comprehensive collection of related works. In this section, we briefly survey existing anti-phishing solutions and list of the related works. One approach is to stop phishing at the email level , since most current phishing attacks use broadcast email (spam) to lure victims to a phishing website . Another approach is to use security toolbars. The phishing filter in IE7 is a toolbar approach with more features such as blocking the user‘s activity with a detected phishing site. A third approach is to visually differentiate the phishing sites from the spoofed legitimate sites. Dynamic Security Skins proposes to use a randomly generated visual hash to customize the browser window or web form elements to indicate the successfully authenticated sites. A fourth approach is two factor authentication, which ensures that the user not only knows a secret but also presents a security token .However, this approach is a server-side solution. Phishing can still happen at sites that do not support two-factor authentication. Sensitive information that is not related to a specific site, *e.g.*, credit card information and SSN, cannot be protected by this approach either. Many industrial anti phishing products use toolbars in Web browsers, but some researchers have shown that security tool bars don‘t effectively prevent phishing attacks. proposed a scheme that utilizes a cryptographic identity verification method that lets remote Web servers prove their identities. However, this proposal requires changes to the entire Web infrastructure (both servers and clients), so it can succeed only if the entire industry supports it. Proposed a tool to model and describe phishing by visualizing and quantifying a given site‘s threat, but this method still wouldn‘t provide an antiphishing solution. Another approach is to employ certification,e.g.(microsoft.com/mscorp/twc/privacy/spam). A recent and particularly promising solution was proposed to combine the technique of standard certificates with a visual indication of correct certification; a site-dependent logo indicating that the certificate was valid would be displayed in a trusted credentials area of the browser. A variant of web credential is to use a database or list published by a trusted party, where known phishing web sites are blacklisted. For example Netcraft antiphishing toolbar http://toolbar.netcraft.com/ prevents phishing attacks by utilizing a centralized blacklist of current phishing URLs. Other Examples include Websense, McAfee‘s anti–phishing filter, Netcraft anti-phishing system, Cloudmark SafetyBar, and Microsoft Phishing Filter **.** The weaknesses of this approach are its poor scalability and its timeliness. Note that phishing sites are cheap and easy to build and their average lifetime is only a few days. APWG provides a solution directory at (Anti-Phishing Working Group) which contains most of the major antiphishing companies in the world. However, an automatic antiphishing method is seldom reported. The typical technologies of antiphishing from the User Interface aspect are done by and . They proposed methods that need Web page creators to follow certain rules to create Web pages, either by adding dynamic skin to Web pages or adding sensitive information location attributes to HTML code. However, it is difficult to convince all Web page creators to follow the rules . The DOM based visual similarity of Web pages is oriented, and the concept of visual approach to phishing detection was first introduced. Through this approach, a phishing Web page can be detected and reported in an automatic way rather than involving too many human efforts. Their method first decomposes the Web pages (in HTML) into salient (visually distinguishable) block regions.

3. FUZZY LOGIC AND DATA MINING

DM is the process of searching through large amounts of data and picking out relevant information. It has been described as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from large data sets.

It is a powerful new technology with great potential to help researchers focus on the most important information in their data archive. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. there are many characteristics and factors that can distinguish the original legitimate website from the forged e-banking phishing website like Spelling errors. The approach is to apply fuzzy logic and RIPPER data mining algorithm to assess phishing email based on the 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.

4. DETECTING AND CLASSIFYING PHISHING EMAILS

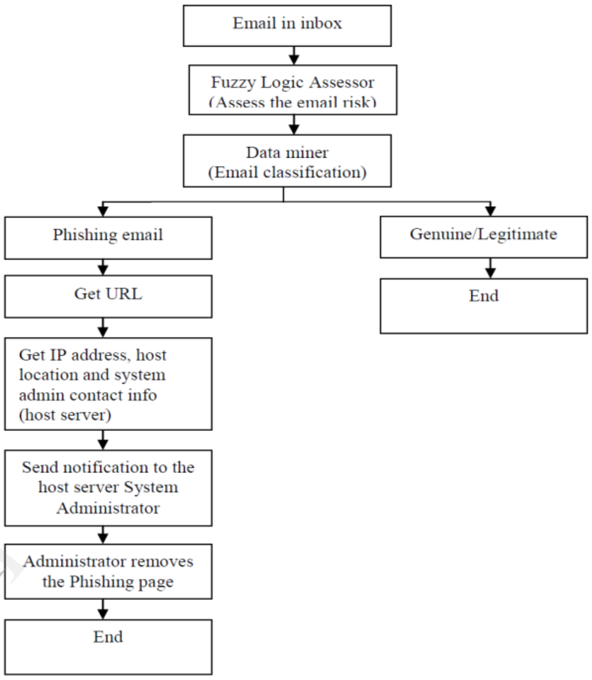
Email in inbox
Fuzzy Logic Assessor
(Assess the email risk)
Data miner
(Email classification)
Phishing email
Genuine/Legitimate
Get URL
End
Get IP address, host location and system admin contact info (host server)
Send notification to the host server System Administrator
Administrator removes the Phishing page
End

Fig 1: Overall system approach

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.

Table 1:Characteristics of phishing emails

Classification approach	**Category/Criteria**	**Component**	**Layer**
Non Content Based Approach	URL	IP URL	1
		Redirect URL	
		Non Matching URL	
		Crawler URL	
		Long URL address	
		URL prefix/suffix	
Content Based Approach	Email Message	Spelling errors	2
		Keywords	
		Embedded Links	

5.MINING USING RIPPER 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. Classification is done using WEKA.

5.1 Algorithm:

Initialize RS = {}, and for each class from the less prevalent one to the more frequent one, DO:

1. Building stage:

Repeat 1.1 and 1.2 until the descrition length (DL) of the ruleset and examples is 64 bits greater than the smallest DL met so far, or there are no positive examples, or the error rate >= 50%.

1.1. Grow Phase

Grow one rule by greedily adding antecedents (or conditions) to the rule until the rule is perfect (i.e. 100% accurate). The procedure tries every possible value of each attribute and selects the condition with highest information gain: p(log(p/t)-log(P/T)).

1.2. Prune phase:

Incrementally prune each rule and allow the pruning of any final sequences of the antecedents;The pruning metric is (p-n)/(p+n) – but it's actually 2p/(p+n) -1, so in this implementation we simply use p/(p+n) (actually (p+1)/(p+n+2), thus if p+n is 0, it's 0.5).

2. Optimization stage:

After generating the initial ruleset {Ri}, generate and prune two variants of each rule Ri from randomized data using procedure 1.1 and 1.2. But one variant is generated from an empty rule while the other is generated by greedily adding antecedents to the original rule. Moreover, the pruning metric used here is (TP+TN)/(P+N).Then the smallest possible DL for each variant and the original rule is computed. The variant with the minimal DL is selected as the final representative of Ri in the ruleset.After all the rules in {Ri} have been examined and if there are still residual positives, more rules are generated based on the residual positives using Building Stage again.

3. Delete the rules from the ruleset that would increase the DL of the whole ruleset if it were in it. and add resultant ruleset to RS.

ENDDO

Rule	IP URL	Redirect URL	Non Matching URL	Crawler URL	Long address URL	URL prefix/ suffix	Output
1	Low	Low	Low	Low	Low	Low	Genuine
2	Low	Low	Low	Low	Low	Moderate	Genuine
3	Moderate	Low	Moderate	Low	Low	Moderate	Suspicious
4	Low	Low	Low	Moderate	Moderate	Moderate	Suspicious
5	Moderate	Moderate	Moderate	High	High	High	Fraud
6	High	High	High	High	Moderate	Moderate	Fraud

5.2 Rule base for layer 1

Rule	Spelling errors	Keywords	Embedded Links	Output
1	Low	Low	Moderate	Genuine
2	Low	Moderate	Moderate	Suspicious
3	High	High	High	Fraud
4	Low	Low	Low	Genuine
5	Moderate	Low	Moderate	Suspicious
6	High	Moderate	Moderate	Fraud

5.3 Rule Base for layer 2

5.4 Locating the Host Server of the Phishing Page

WHOIS is a protocol used to find information about networks, domains and hosts. The WHOIS query is used to locate the host server of a phishing page. WHOIS is a query/response protocol that is widely used for querying an official database. The WHOIS database contains IP addresses, autonomous system numbers, organizations or customers that are associated with these resources, and related Points of Contact on the Internet . A WHOIS search will provide information regarding a domain name, such as example.com. It may include information, such as domain ownership, where and when registered, expiration date, and the name servers assigned to the domain. The system runs the WHOIS query on the URL that is contained in the Phishing email.Upon receiving the notification of the phishing page's existence on the host server,the hosting administrator will then test the legitimacy of the phishing link and its validity. Once the Administrator confirms the phishing page, the infected or hacked website will be shut down immediately to protect Internet users from further phishing. The host Administrator then notifies the website owner about the existence of the phishing page within their website. As soon as the phishing page is removed, if no notification has been sent, the proposed system will periodically check for evidence that it has been removed. This technique assumes that website owner and host Administrator are absolutely unaware of the presence of the phishing page within their website or server until our technique notifies them. This means Phishers are taking control of the legitimate website to upload their phishing page.

5.5 Removal of Phishing page:

Upon receiving the notification of the phishing page's existence on the host server,,the hosting administrator will then test the legitimacy of the phishing link and its validity. Once the Administrator confirms the phishing page, the infected or hacked website will be shut down immediately to protect Internet users from further phishing. The host Administrator then notifies the website owner about the existence of the phishing page within their website. As soon as the phishing page is removed, if no notification has been sent, the proposed system will periodically check for evidence that it has been removed. This technique assumes that website owner and host Administrator are absolutely unaware of the presence of the phishing page within their website or server until our technique notifies them. This means Phishers are taking control of the legitimate website to upload their phishing page.

6. RESULTS

100 websites from Phishtank.com were considered for testing purpose. 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).

Table 4:Results from WEKA

Validation mode	10 folds cross validation
Attributes	URL Domain
	Email Content
Number of Rules	12
Correctly classified	85.4%
Incorrectly Classified	14.6%
No. of samples	100

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.

7. CONCLUSIONS AND FUTURE WORK

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. The initial objective is to assess the risk of the email in the archive data using fuzzy logic and the RIPPER classification algorithm. Several characteristics were identified and major rules that were determined along the study were used in the fuzzy rule engine. The results showed that the RIPPER algorithm achieved 85.4% for correctly classified Phishing emails and 14.6% for wrongly classified Phishing emails. The phishing page removal success rate is 81.81%.

REFERENCES

[1] WholeSecurity Web Caller-ID, www.wholesecurity.com

[2] Anti-Phishing Working Group. Phishing Activity Trends Report, http://antiphishing.org/reports/apwg_report_sep2007_final.pdf. September 2007.

[3] B. Adida, S. Hohenberger and R. Rivest, ―Lightweight Encryption for Email,‖ USENIX Steps to Reducing Unwanted Traffic on the Internet Workshop (SRUTI), 2005.

[4] 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.

[5] R. Dhamija and J.D. Tygar, ―The Battle against Phishing: Dynamic Security Skins,‖ Proc. Symp. Usable Privacy and Security, 2005.

[6] FDIC., ―Putting an End to Account-Hijacking Identity Theft,‖ http://www.fdic.gov/consumers/consumer/idtheftstudy/identity_theft.pdf 2004.

[7] A. Y. Fu, L. Wenyin and X. Deng, ― Detecting Phishing Web Pages with Visual Similarity Assessment Based on Earth Mover‗s Distance (EMD) ‖ IEEE transactions on dependable and secure computing, vol. 3, no. 4, 2006.

[8] A. Herzberg and A. Gbara, ―Protecting Naive Web Users,‖ Draft of July 18, 2004.

[9] C. Y. Ho, B. W. Ling and J. D. Reiss, "Fuzzy Impulsive Control of High-Order Interpolative Low-Pass Sigma–Delta Modulators," IEEE Transactions on Circuits and Systems—I: Regular Papers, Vol. 53, No 10, October 2006.

[10] L. James**,** ―Phishing Exposed,‖ Tech Target Article sponsored by: Sunbelt software, searchexchange.com, 2006.

[11] M. Liu, D. Chen and C. Wu. "The continuity of Mamdani method," International Conference on Machine Learning and Cybernetics, Page(s): 1680 - 1682 vol.3, 2002.

[12] W. Liu, G. Huang, X. Liu, M. Zhang, and X. Deng, ―Phishing Web Page
Detection,‖ Proc. Eighth Int‘l Conf. Documents Analysis and Recognition, pp. 560-564, 2005.

[13] W. Liu, X. Deng, G. Huang and A. Y. Fu, ―An Antiphishing Strategy Based on Visual Similarity Assessment,‖ Published by the IEEE Computer Society 1089-7801/06 IEEE , INTERNET COMPUTING IEEE, 2006.

[14] Microsoft Corp, ―Microsoft Phishing Filter: A New Approach to
Building Trust in E-Commerce Content,‖ White Paper, 2005.
[15] S. Olsen, ―AOL tests caller ID for e-mail,‖ CNET News.com, January
22, 2004.