A NOVEL APPROACH FOR PHISHING WEBSITE DETECTION USING RULE MINING AND MACHINE LEARNING TECHNIQUE

BINAL MASOT[a1], RIDDHI KOTAK[b] AND MITTAL JOISER[c]

ABSTRACT

In last few years, phishing is a major problem of web or internet because the internet has become a crucial part of our daily life activity like reading a newspaper, online shopping, online payment etc. Hence internet users may be unsafe to a typical types of web attacks which may induce loss of the financial, personal information, brand name reputation customer trust from online transaction. There for the phishing detection necessary. There is no conclusive solution to detect phishing. In this paper we present main two core parts 1) To details investigation on phishing circumstance and 2) proposed spearhead framework to detect phishing attack. Our proposed framework work on combine algorithm of rule mining and machine learning.in this first rule mining algorithm is applied after the result of it machine learning algorithm is applied so we can get better accuracy.

KEYWORDS: Data Mining, Feature Extraction, Legitimate, Machine Learning, Phishing

Now a day the most profitable fraud is 'identity thievery' means that to take users personal information. The word 'phishing' is derived from the word "fishing + phreaking", fishing means use bait to induce the target and phreaking. The word "phishing" was first used in 1996 over the internet by a group of hackers who stole America online (AOL) accounts. By tricking unaware AOL users into disclosing their passwords [Gupta et. al., 2016]. The main aim of phishing attack is to steal private delicate information such as usernames, passwords, credit card details, confidential information, bank information, employment details, financial record, and electricity bills and so on. Last few years phishing quickly spread posing a real threat to universal security.

Website phishing refers to the form of web threat that indirectly get information of victim like personal data, credential information. Phisher will create a replica of legitimate website so user cannot identify directly. The different technique of Phishing by send email of fake site URL hyperlink, instant message, website and SMS.

In this paper, include overview of phishing attacks, set of features used for detection of phishing, performance metrics to find accuracy. We also provide proposed solution that can detect phishing attacks.

NARRATIVE

The history of the phishing start from the 1996, day by day rate of the attack is increase. Table 1 shows growth rate of phishing starts from 1996 to till now according to RSA online fraud report [PGU&PPA, 2007, APWG, 2014].

Table I: Evaluation of Phishing during 1996-2016

Year	Occurrence
1996	"Phishing" word first used
1997	Declared a new threat called "Phishing"
1998	Starting medium of attackers was message and newsgroups.
1999	Using the Email system for the phishing attack.
2000	Phishers used key loggers type attack for getting login details
2001	Used URL to direct user to making a fake site
2002	Used screen loggers attack
2003	Used IM and IRC
2004	Evolvement of "pharming"
2005	First used spear phishing word
2006	First phishing attack over VoIP
2007	Become phishing scams more than $3 billion
2008	Increased 39.8% than previous year
2009	SHS blocked phishing attacks Impersonating 1079 different organizations
2010	Facebook attracted more phishing attacks compare to Google and IRS
2011	Web Hacking Incident Database(WHID)
2012	Identified 6 million unique malware sample
2013	69 Countries scam over Red October Operation
2014	Used of IOT 7,50,000 malicious emails sent
2015	Spear phishing reached
2016	Unsolicited emails containing malicious attachment

PERIOD OF EXISTENCE

Any attack have some period to existence.

[1]Corresponding author

Phishing attack have some step or life cycle to attack on user. The following stage are involved in phishing life cycle as shown in Fig1.

Step 1 Analysis and Environment setup

This is the first step or initialization step of the phishing. In this step the attackers analysed the organization and which types of network it's used. Then set the environment. e.g., make a replica of legitimate website, which may redirect the victim to some fraud web page.

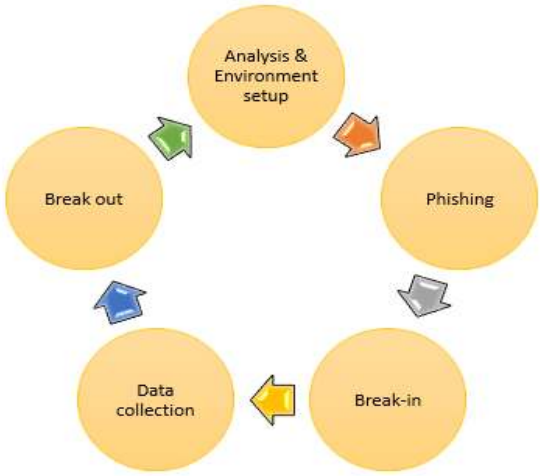


Figure 1: Phishing Life Cycle

Step 2 Phishing

After successful of setup the next step is send to the fraud mail or link a spoofed website, e.g, ask user to update some sensitive information urgently by clicking on some malicious link. Another example is linked with phishy URL instead of legitimate, e.g., www.faceb00k.com.

Step 3 Break-in

As soon as the victim open fraud link, a malware is installed on the system which allows the attacker to intrude the system and change it configuration or access rights.

Step 4 Data collection

Once the attackers get access to the victim system, the required data and account detail are extracted. Phisher use rootkits to hide their malwares.

Step 5 Break-out

After getting the required information the phisher remove all the link and website.it is also observed that they track the degree of success of their attack for refining future attack.

LITERATURE REVIEW

Gupta et al., 2016 propose the survey on fighting against phishing attack. They give the various challenges and available solution. Jeeva and Rajsingh, 2016 propose an approach is based on the association rule mining to detect phishing URL. This approach in two phase in first phase they search URL and in second phase they extract the features. The result show that the proposed method achieved overall 93% accuracy.

Gowtham et. al., 2014 proposed an anti-phishing technique using target domain identification in this they take a groups the domain from hyperlinks having direct or indirect association with the given suspicious webpage. The result show that the proposed method achieved 99.65% accuracy on google.com search engine, 99.6% on aol.com search engine, 99.55% on hotbot.com search engine, 99.45% on bing.com search engine.

Khonji et al., 2011 proposed the technique of phishing detection to reduce the rate of false positive ratio. The main aim of this paper is to extract the domain name from the victim URL and compare the page rank of this extracted domain name with actual domain name. if not same then domain name will be reported as phishing.

Shrestha et al., 2015 proposed a multi label feature classification algorithm to classify whether a website is phishing or legitimate. In this text based feature used to implementation extracts visual feature from the screenshot of a phishing website and text from its html source code. This technique 30 times faster than existing state of the art system in phishing website classification problem.

PERFORMANCE EVALUATION MATRICES

The main aim of most classifiers is to perform binary classification, i.e., phishing or legitimate. There are main four possibility exits to find the performance. These four possibility are True Positive, True Negative, False Positive and False Negative.

Assume that N_H denotes the total number of ham email and N_P denotes the total number of phishing email. If ($n_h \rightarrow H$) denotes ham message, then ($n_p \rightarrow H$) denotes phishing emails classified as ham ($n_h \rightarrow P$) denotes ham mails classified as phishing and ($n_p \rightarrow P$)

denotes phishing emails classified as phishing. The evaluation metrics used in this case are [Husna et. al., 2008 & Toolan and Carthy, 2009]:

1) True Positive (TP):- Ratio of the number of phishing website is identified correctly as:

$$TP=\frac{np\rightarrow P}{Np}$$

2) True Negative (TN): Ratio of the number of ham website identified correctly as:

$$TN=\frac{nh\rightarrow H}{Nh}$$

3) False positive (FP): Ratio of the number of ham website classified as phishing, as:

$$FP=\frac{nh\rightarrow P}{Nh}$$

4) False negative (FN): Ratio denoting the number of phishing website classified as ham, as:

$$FN=\frac{np\rightarrow H}{Np}$$

5) Precision (P): Measures the rate of phishing website which are identified correctly as the website detected as phishing:

$$P=\frac{TP}{TP+FP}$$

6) Recall (R): Measures the rate of phishing website which are identified correctly as existing phishing website:

$$R=\frac{TP}{TP+FN}$$

7) F_1 Score: This is the harmonic mean of Precision and Recall:

$$F_1=\frac{2PR}{P+R}$$

8) Accuracy (ACC): Measures overall correctly identified website:

$$ACC=\frac{TP+TN}{TP+TN+FP+FN}$$

FEATURES USED FOR IDENTIFICATION OF PHISHING WEBSITE

The importance of features is to help the algorithm to give an accurate result. Toolan and Carthy, 2009 studied the utility of about 40 such features we have categorized URL features used for detection of phishing website as follow:

IP address

In general the legitimate site have a domain name. If the presence of the IP address in the URL instead of using the domain name of the website that indicate someone is trying to access your personal information. An IP address is like http:// 91.121.10.211/~chems/websce/verify. Sometime an IP address is transfer into hexadecimal like http://0x58.0xCC.0xCA.0x62.

Rule:-

If (IP address exists in URL then) ⟶ phishing

Else⟶ non-phishing

Length of URL

URL of the website consist three element network protocol, host name and path. For a given URL extracted the total length of the URL. If the length of URL is greater than 40 character then the site is phishing otherwise legitimate. i.e.http:// face book .com.bugs3.com/login/Secured_Relogin/index1. html.

Rule:-

If (host name)> 40 character⟶ phishing

Else ⟶ non-phishing

Number of dots in URL

This feature verify the presence of the dot in host name of the URL. Phishing site usually puts extra dots in URL to make users believe that they are legit page. i.e.http://www.Facebook.pcriot .com/ login.php.

Rule:-

If (Number of dots)> 4⟶phishing

Else⟶ non-phishing

Number of suspicious URL

@, _ ,-- is the suspicious characters, if in URL suspicious character present then that website is phishing. The ‘‘@’’ symbol leads the browser to ignore everything suffix it and redirects the user to the link typed after @ symbol. i.e. http://faceebook-com.bugs3.com/ login/Secured_Re-login/index1.html.

Rule:-

If (URL has suspicious) phishing

Else ⟶ non-phishing

Number of slashes in URL

Additional slashes in URL such a technique to make a mimic URL look legitimate. If the URL contain 5 or more than 5 then the site is phishing. i.e. http://faceebook-com. bugs3. Com /login /Secured_Re-login/index1.html.

Rule:-
If (slash in URL)>=5 ⟶ phishing
Else⟶non-phishing

"WHOIS" lookup

WHOIS is a protocol which used to fetch the customer detail of the registered website from the database. Legitimate website always stored in WHOIS data base.

Rule:-
If (not in WHOIS database) ⟶phishing
Else ⟶ non-phishing

Length of host name in URL

URL string consist three element network protocol, host name and path. For a given URL extracted the length of the host name. If the length of host name is greater than 25 character then the site is phishing otherwise legitimate.

Rule:-

If (host name)> 25 character ⟶ phishing

Else⟶ non-phishing

Age of domain

It can be extracted from WHOIS database. A PHP script was created to connect to WHOIS database. If the domain age is less than one year then it classified as a 'phishing', else if the domain age is more than one and less than 2 year then it classified as suspicious, else it is legitimate.

Rule:-
If (age of domain) <1 year ⟶phishing
Else if (age of domain) <2 year ⟶suspicious
Else⟶ non-phishing

Unicode in URL

In URL consist the unique number for every character. i.e. http://www.paypa1.com. In this URL 1 is represent the l

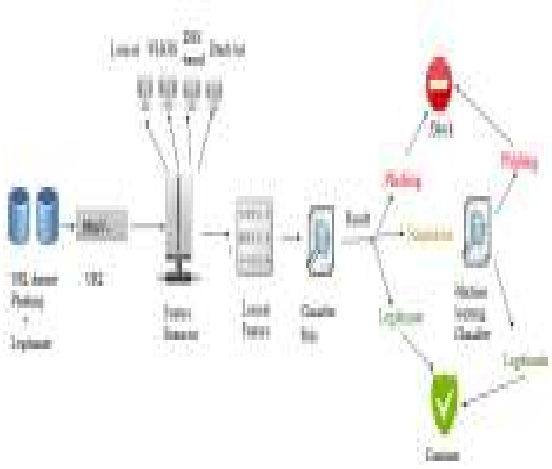
Rule:-

If Unicode ⟶ phishing

Else⟶ non-phishing

PROPOSED WORK

The sources of phishing attacks are mostly from email, websites and malware. The links (URL) provided in phishing emails draws user into entering phishing website. In website based phishing, website is replica of trusted website users into revealing sensitive information. There are several technique to detect phishing. All Applied techniques contains mixture of features like content based, lexical based, body based and so on. In our proposed system only used the URL based features. Benefit is used to URL features is if we used content based or body based we classify the whole



source code of the webpage so its time consuming.

Figure 2: Proposed frame work

In this proposed system dataset is taken from the different data source. In our system the dataset is a mixing of phishing and legitimate URL. For phishing data we collect from the Phish Tank API data source and for non-phishing data we collect from the Alexa Database.

Our system works on combination of rule mining [3] and Machine Learning [4] algorithm. First using if-else mining to classify the URL in three form phishing, legitimate and suspicious. Then take the suspicious URL and applied the Machine Learning algorithm to classify the Suspicious URL is phishing or

legitimate. So overall we classify the all the URL in two form phishing and legitimate.

CONCLUSION

This research present a details of phishing attack. For phishing detection we analyzed the URL features using the if-else rules it is hybrid with machine learning technique to solve the suspicious URL problem. Analyzed features are more sensible to phishing detection URL.so our proposed work easily find the phishing website and if find the phishing URL then its puts in blacklist automatically prevent.

REFERENCES

Gupta B. B., Tewari A., Jain A.K. and Agrawal D.P., 2016. "Fighting against phishing attacks: state of the art and future challenges" Neural Computing and Applications springer, pp. 1–26.

The Phishing Guide Understanding & Preventing Phishing Attacks By: Gunter Ollmann, Director of Security Strategy, IBM Internet Security Systems, 2007.

Jeeva and Rajsingh, 2016. "Intelligent phishing url detection using association rule mining" Human-Centric Computing Information Sciences (2016) springer, pp. 1-19.

Huang H., Qian L. and Wang Y., 2012. "A SVM based technique to detect Phishing URLs", Information Technology Journal, **11**(7): 921-925.

Gowtham R., Kumar K.S.S. and Krishnamurthi I., 2014. "An efficacious method for detecting phishing webpage through Target Domain Identification" Decision Support Systems (2014) Elsevier, **61**:12–22.

Dhamija R. and Tygar J.D., 2006. "Hearst MA (2006) Why phishing works,'' in proceedings of the conference on human factors in computing systems (CHI). ACM, Montre´al, Que´bec, Canada, pp 581–590.

Anti-Phishing Working Group (APWG) (2014) Phishing activity trends report—first quarter 2014. http://antiphishing.org/reports/apwgtrendsreportq12014.pdf. Accessed Sept 2014.

Anti-Phishing Working Group (APWG) (2014) Phishing activity trends report—fourth quarter 2013. http://antiphishing.org/reports/apwgtrendsreportq42013.pdf. Accessed Sept 2014.

Anti-Phishing Working Group (APWG) (2014) Phishing activity trends report—second quarter 2013. http://antiphishing.org/reports/apwgtrendsreportq22013.pdf. Accessed Sept 2014.

Shrestha N., Kharel R.K., Britt J. and Hasan R., 2015. "High Performance classification of phishin URLs using a multimodel Approach with MapReduce", IEEE world congress on date of conference, pp. 206-212

Khonji M., Jones A. and Iragi Y., 2011. "A novel Phishing Classification based on URL Features", IEEE GCC Conference and exihibition(GCC),Dubai,United Arab Emirates, pp.19-22.

Abdelhamid N., Ayesh A. and Thabtah F., 2014. "Phishing detection based associative classification data mining" Science- Direct, pp.5948–5959.

Agrawal R., Imielinski T. and Swami A., 1993. "Mining association rules between sets of items in large databases" ACMSIGMOD, pp.207–216.

Aburrous M., Hossain M.A., Dahal K. and Thabtah F., 2010. "Predicting phishing websites using classification mining techniques with experimental case studies" Seventh international conference on information technology. IEEE Conference, Las Vegas, Nevada, USA, 2010, pp 176–181.

Husna H., Phithakkitnukoon S., Palla S. and Dantu R., 2008. "Behavior analysis of spam botnets". In: Communication systems software and middleware and workshops, 2008. COMSWARE 2008. 3rd International Conference, Bangalore, India, 2008, pp 246 253.

Toolan F. and Carthy J., 2009. "Phishing detection using classifier ensembles". In: eCrime researchers summit, IEEE conference Tacoma, WA, USA, 2009, pp 1–9.