IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Detection of Phishing Websites using an Efficient Deep Learning Framework

Ashwini R[1] , Gayathri R[2] , Kaviya B[3] , Mohammad Bilal N[4] , MynthuriIswarya S[5]

Assistant Professor, Department Of C.S.E, Jansons Institute Of Technology, Coimbatore,India[1]

UG Students , Department Of C.S.E, Jansons Institute Of Technology, Coimbatore,India[2-5]

ABSTRACT

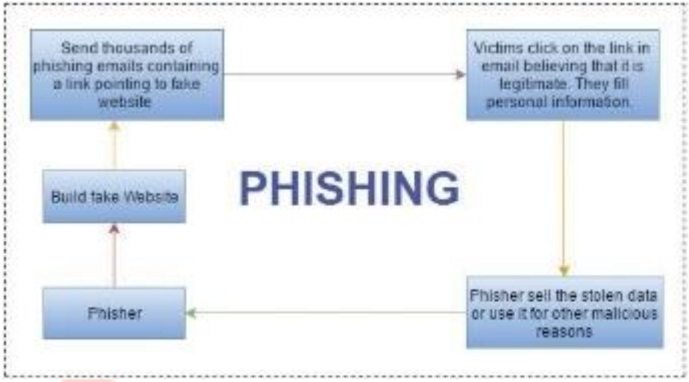
There are number of users who purchase products online and make payment through various websites. There are multiple websites who ask user to provide sensitive data such as username, password or credit card details etc. often for malicious reasons. This type of websites is known as phishing website. Phishing attacks are becoming more common and sophisticated, putting Internet users at risk. While these assaults have shown robust to a wide number of countermeasures presented by academia, business, and research groups, machine learning algorithms look to be a viable option for discriminating between phishing and authentic websites. Existing machine learning algorithms for phishing detection have three major drawbacks. The first issue is that there is neither a framework for extracting features and maintaining the dataset up to date, nor an updated collection of phishing and genuine websites. The second point of concern is the vast number of features employed, as well as the absence of supporting evidence for the characteristics used to train the machine learning classifier. The final point of concern is the sort of datasets utilised in the research, which appear to be unwittingly skewed in terms of URL or content-based attributes. The development of our open-source and extendable system to extract features and produce up-to-date phishing dataset is described in this thesis. We integrated 29 distinct characteristics into our framework, dubbed Fresh-Phish, to determine whether a particular website is authentic or phishing. We constructed a dataset of 6,000 websites with these qualities, 3,000 of which were malicious and 3,000 of which were legitimate, and evaluated our technique using 26 features published in previous work and three novel features. We concentrate on this aspect of phishing websites and design features that investigate the domain name's relationship to the website's key elements. Our study varies from the previous state-of-the-art in that our feature set assures that a dataset has low or no bias. On a sample dataset, our learning model achieves a true positive rate of 98 percent and a classification accuracy of 97 percent using only seven features. Our per data instance processing and classification is 4 times quicker for authentic websites and 10 times faster for phishing websites when compared to state-of-the-art work. We also show the drawbacks of utilising URL-based characteristics, since they are likely to be skewed towards dataset acquisition and consumption.

INTRODUCTION

Phishing is a form of fraudulent attack where the attacker tries to gain sensitive information by posing as a reputable source. In a typical phishing attack, a victim opens a compromised link that poses as a credible website. The victim is then asked to enter their credentials, but since it is a “fake” website, the sensitive information is routed to the hacker and the victim gets ”‘hacked.”Social engineering attack is a common security threat usedto reveal private and confidential information by simply tricking the users without being detected.The main purpose of this attack is to gain sensitive information such as username, password and account numbers. According to, phishing or web spoofing technique is one example of social engineering attack. Phishing attack may appear in many types of communication forms such as messaging, SMS,VOI Pand fraud steremails. Users commonly have many user accounts on various websites including socialnetwork, email and also accounts for banking. Therefore, the innocent web users are the most vulnerable targets towards this attack since the fact that most people are unaware of their valuable information,which helps to make this attack successful.

Typically phishing attack exploits the social engineering to lure the victim through sending as poofed link by redirecting the victim to a fake webpage.The spoofed link is placed on the popular web pages or sent via email to the victim. The fake webpage is created similar to the legitimate webpage.Thus,rather than directing the victim request to the real webserver,it will be directed to the attacker server.

The current solutions of antivirus, firewall and designatedsoftware do not fully prevent the web spoofing attack. Theimplementation of Secure Socket Layer (SSL) and digital certificate (DA) also does not protect the web user against such attack.In web spoofing attack,the attacker diverts the request to fake webserver.In fact,acertain type of SSL and DA can be forged while everything appears to belegitimate.According to, secure browsing connection does virtually nothing to protect the users especially from the attackers that have knowledge on how the“secure” connections actually work. This paper develops an anti-web spoofingsolution based on inspecting the URLs of fake web pages.This solution developed series of steps to check characteristics of websites Uniform Resources Locators(URLs). URLs of a phishing webpage typically have someunique characteristics that make it different from the URL soft helegitimate webpage.Thus,URL isused in this paper to determine the location of the resource in computer networks.


Send thousands of phishing emails containing a link pointing to fake website
Victims click on the link in email believing that it is legitimate. They fill personal information.
PHISHING
Build fake Website
Phisher
Phisher sell the stolen data or use it for other malicious reasons

Flow of general phishing attack

LITERATURESURVEY

In [1] JAINMAO, WENQIANTIAN and ZHENKAI LIANG has proposed a system which detect the phishing using pagecomponentsimilaritywhichanalyzesURLtokens toincreaseprediction accuracy phishing pages typically keep its CSSstyle similar to their targetpages.Based on the observation,a straightforward approach to detect phishing pages is to compare all CSS rules of two webpages,It prototyped Phishing-Alarm as an extension to the Google Chrome browser and demonstrated its effectiveness in evaluation using real-world phishing samples.

ZOU FUTAI, PEI BEI and PAN LI [2] Uses GraphMiningtechnique for web Phishing Detection. It can detect some potential phishing which can't be detected by URL analysis. It utilize the visiting relation between user and website. To get dataset from the real traffic of a LargeISP.After anonymizing these data, they have cleansing dataset and each record includes eight fields: User node number (AD),User SRC IP(SRC-IP) access time (TS), Visiting URL pool.Therefore,we build the visiting relation graph with AD and URL,calledAD-URL Graph and the Phishing website is detected through the Mutual behavior of the graph

In [3] NICK WILLIAMS and SHUJUN LI proposed a system which analysis ACT-Rcognitive behavior architecture model. Simulate the cognitive processes involved in judging the validity of are presentative webpage based primarily around the characteristics of the HTTPS pad lock security indicator.ACT-Rpossesses strong capabilities which map well on to the phishing usecase and that further work to morefully represent the range of human security knowledge and behaviors in an ACT-Rmodel could lead to improved in sights into how best to combine technical and human defenses to reduce the risk to users from phishing attacks

XINMEICHOO,KANGLENGCHIEWandNADIANATRA MUSA[4] this system is based on utilizing support vector machine to perform the classification.This method will extract and form the feature set for a webpage.It uses a SVMmachine as a classifier which has two phase training phase and testing phase during training phase it extracts feature set and while testing it predict the website is legitimate or a phishing.

In[5] GIOVANNI ARMANO, SAMUEL MARCHAL and N.ASOKAN proposed a use of add on in the browser which is Real-TimeClient-Side Phishing Prevention It uses information extracted from website visited by the user to detect if it is a phish and warn the user. It also determines the target of the phish and offers to redirect the user there.Awarning message is displayed in the foreground while the background displays the phishing webpage darkened by ablack semi-transparent layer preventing interactions withthewebsite.

TRUPATIKUMBHARE and SANTOSHCHOBE [6]have discussed various Association Rule Mining Algorithm.Association rule learning searches for relationships among variables. Various Association algorithm discussed are AIS algorithm, SETM algorithm,Apriori algorithm,Aprioritid algorithm,Apriorihybrid algorithm,and FP-growth algorithm.

In[7]S.NEELAMEGAM and DR.E.RAMARAJ discussed various Classification Algorithm used in datamining.Data Classification is a data mining technique used to predict group membership for data instances Various Classification Algorithm discussed are decision tree, Bayesian networks,k-nearest neighbor classifier, Neural Network, Support vector machine.

VARSHARANIRAMDAS,V.Y.KULKARNIandR.A.RANE[8] proposed a system to detect a phishing website using Novel Algorithm .This detection algorithm can find out the maximum number of phishing URLs because it executes multiple tests such as Blacklistsearch Test, Alexa ranking test, and different URL features test.But this solution is effective only for HTTPURLs.

In[9]JUNHU,YUCHUNJI HANBINGYAN this method to detect Phishing website is based on analysis of legitimate website server login information. Everytime a victim opens the phishing website, the phishing website will refer to the legal website by asking for resources. Then, there will be alog,which is recorded by the legitimate website server and from this logs Phishing site can be Detected

SAMUELNARCHAL,GIOVANNI ARMANO and NIDHI SINGH[10] propose a application Off-the-Hook application for detection of phishing website. Off-the-Hook, exhibits several not able properties including high accuracy, brand-independence and good language-independence, speed of decision, resilience to dynamic phish and resilience to evolution in phishing techniques.

EXISTING SYSTEM

Existing system use different and multiple substitution or permutation processes to shuffle and distort the pixels of an image.NIST based randomness results as compared to existing methods.histogram consistency analysis and its variance(HCAV)are used.

PROPOSEDSYSTEM

It describes the proposed model of phishing attack detection. The proposed model focuses on identifying the phishing attack based on checking phishing websites features, Blacklist and WHOIS database. According to few selected features can be used to differentiate between legitimate and spoofed web pages. These selected features are many such as URLs, domain identity, security & encryption,sourcecode,page style and contents,webaddress barand social human factor.This study focuses only on URLs and domain name features. Features of URLs and domain names are checked using several criteria such as IPAddress,longURLaddress,adding a prefix or suffix,redirecting using the symbol“//”, and URLs having the symbol“@”.These features are inspected using a set of rules in order to distinguish URLs of phishing webpages from the URLs of legitimate websites.

A.URLbased

UsingtheIPAddress

IfanIPaddressisusedasanalternativeofthedomain namein the URL, such as “http://125.98.3.123/fake.html”, userscan be sure that someone is trying to steal their personalsensitive information. Sometimes, the IP address is eventransformedintohexadecimalcodeasshowninthefollowinglink

“http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.

html”.Rule:IF The Domain arthasan IP Address

→ Otherwise→ Legitimate

LongURLtoHidetheSuspiciousPart

Phishers can use long URL to hide the doubtful part in the address bar.
For example:

http://federmacedoadv.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/?cmd=_home&amp;dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@phishing.website.html
Toensuretheaccuracyofourstudy,wecalculatedthelengthofURLsinthedatasetandproducedanaverageURLlength.The results showed that if the length of the URL is greaterthanorequal54charactersthentheURLclassifiedasphishing. By reviewing our dataset we were able to find1220URLslengthsequalsto54ormorewhichconstitute48.8%ofthetotal datasetsize.

Rule: IF URL length is

≤75→legitimate

otherwise→Phishing

We have been able to update this feature rule by using amethod based on frequency and thus improving upon its accuracy.

AddingPrefixorSuffixSeparatedby(-)totheDomain

The dash symbol is rarely used in legitimate URLs.Phisherstend to add prefixes or suffixes separated by(-)to the domain name so that users feel that they are dealing with a legitimate webpage.Forexample http://www.Confirme-paypal.com/.

Rule: IF Domain Name Part

Includes(-)Symbol →

Phishing

Otherwise→Legitimate

SubmittingInformationtoEmail

Web form allows a user to submit his personal sensitive information that is directed to some server for processing.Aphishermightredirecttheuser’s informationtohispersonalemail. To that end, a server-side script language might beused such as “mail()” function in PHP. One more client-side function that might be used for this purpose is the“mailto:”function.

Rule: IF Using ""mail()\" or \"mailto:\" Function to SubmitUser Information"→ Phishing

Otherwise→Legitimate

UsingPop-upWindow

It is un usualt of in dalegitimate website asking users to submit their personal information through a pop-upwindow. On the other hand, this feature has been used in some legitimate websites and its main goal is to warn users about fraudulent activities or broadcast a welcome announcement, though no personal information was asked to be filled in through these pop-up windows.

Rule:IFPopup Window

Contains Text Fields

Phishing Otherwise →

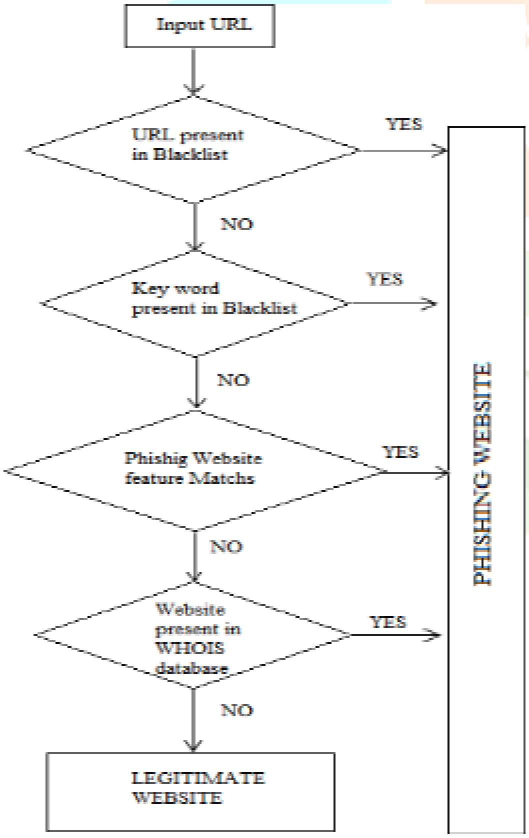
Legitimate

B.Blacklistbased

A Blacklist is created in the proposed model in which the website detected as phishing is saved for the future use at keep a track record and data of the phishing website this can be useful in analyzing the phishing website to increase the efficiency of the system.

C.WHOIS Database

The life of phishing site is very short, therefore; this DNS information may not be available after sometime.If the DNS record is not available anywhere then the website is phishing. If the domain name of the suspicious webpage is not match with the WHOIS database record, then webpage considers as phishing.


Input URL
URL present in Blacklist
YES
NO
Key word present in Blacklist
YES
NO
Phishig Website feature Matchs
YES
NO
Website present in WHOIS database
YES
NO
LEGITIMATE WEBSITE
PHISHING WEBSITE

MODULES

- **Route Recommendation:**

With the availability of user-generated trajectory information, routerecommendation has received much attention from the research community.

- Modeling the Observable Cost with **Black box**
- **Simple frequency-based estimation method** will suffer from data sparsely in large search space even with first-order Markov assumption. In this case, the computed observable cost will not be reliable to be used.
- **Inter-Trajectory Attention**
  The information from a single trajectory is usually limited. In order to capture overall moving patterns for a specific user, we further consider in corporating historical trajectories generated by the user

CONCLUSION

The most important way to protect the user from phishingattack is the education awareness. Internet users must beaware of all security tips which are given by experts. This application can be used by many E-commerce enterprises in order to make the whole transaction process secure. Every user should also be trained not to blindly follow the links to websites where they have to enter their sensitive information. It is essential to check the URL before entering the website. In Future System can upgrade to automatic Detect the webpage and the compatibility of the Application with the web browser. Additional work also can be done by adding some other characteristics to distinguishing the fake web pages from the legitimate web pages. Phish Checker application also can be upgraded into the web phone application in detecting phishing on them obile platform.

REFERENCES

[1]jian mao1,wenqian tian1, pei li1, tao wei2, and zhenkailiang 3 phishing-alarm:robustand efficient phishing detection via page component similarity.

[2]zou futai, gang yuxiang, pei bei, pan li, li linsen web phishing detection based on graph mining.

[3]nickwilliams,shujunlisimulatinghumandetectionofphishingwebsites:aninvestigationintotheapplicabilityofact-rcognitive behavior architecture model.

[4]xin mei choo, kang leng chiew,dayang h ananiabang ibrahim, nadianatra musa, san nah sze, weikingtiongfeature-based phishing detection technique.

[5]giovanniarmano,samuelmarchalandnasokanreal-timeclient-side phishing prevention add-on.

[6]truptia.kumbhare and prof. santoshv. chobe an overview of association rule mining algorithms.

[7]s.neelamegam,dr.e.ramarajclassificationalgorithmindatamining:anoverview

[8]varsharani ramdas hawanna, v. y. kulkarni and r. a.raneanovel algorithmto detect phish in gurls.

[9]junhu,xiangzhuzhang,yuchunji,hanbingyan,liding,jia li and huiming meng detecting phishing websites based on the study of the financial industry webserver logs.

[10]w. d. yu, s. nargundkar, n. tiruthani, "phish catch – aphishing detection tool",33rdannual ieee international on computer software and application sconference 2009. compsac'09,pp.451-456,2009.

[11]A. Y. Fu, L. Wenyin, X. Deng, "Detecting phishing webpages with visual similarity assessment based on earth mover's distance (emd)",IEEE Trans. Dependable Secur. Comput., vol.3, o.4, pp. 301-311, Oct. 2006.

[12]G.Liu,B.Qiu,L.Wenyin,"Automaticdetectionofphishing target from phishing webpage", Pattern Recognition(ICPR) 010 20th International Conference on, pp.4153-4156,aug.2010