ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 5, March 2017

DETECTION OF PHISHING USING DATA MINING

K.Dhanalakshmi[1], T.Sumithra[2], K.Leelarani[3]

UG Student[12], AssistantProfessor[3]

[123]*Department of Computer Science and Engineering*,

Kamaraj College of Engineering and Technology, Affiliated to Anna University,

Kallikudi, Virudhunagar - 626001,Tamilnadu, India

[1]dhanalakshmisugirtha@gmail.com[2]sumitamil26may@gmail.com[3]leelarani90@gmail.com

Abstract:

Phishing websites impersonate legitimate counterparts to lure users into visiting their websites. Once users visit a phishing website then the phishing website may steal users' private information or cause drive-by downloads. To detect a phishing website, human experts compare the claimed identity of a website with features in the website. For example, human experts often compare the domain name in the URL against the claimed identity. Most legitimate websites have domain names that match their identities, while phishing websites usually have less relevance between their domain names and their claimed (fake) identities. Once the URL is identified as phish then the user is not allowed to visit the Phishing Website. This will give awareness about the Phishing attack to the user. If the given URL is labelled as a non-phishing website, then the users are allowed to continue their action.

Keywords

Phishing, anti- phishing, phisher, confidential information, non-phishing

1. INTRODUCTION

Phishing is a new word produced from 'fishing', it refers to the act that the attacker allure users to visit a faked Web site by sending them faked e-mails (or instant messages), and stealthily get victim's personal information such as user name, password, and national security ID, etc. This information then can be used for future target advertisements or even identity theft attacks (e.g., transfer money from victims' bank account). The frequently used attack method is to send e-mails to potential victims, which seemed to be sent by banks, online organizations, or ISPs.

In these e-mails, they will make up some causes, e.g. the password of your credit card had been mis-entered for many times, or they are providing upgrading services, to allure you visit their Web site to conform or modify your account number and password through the hyperlink provided in the e-mail. If You input the account number and password, the attackers then successfully collect the information at the server side, and is able to perform their next step actions with that information (e.g., withdraw money out from your account).Phishing itself is not a new concept, but it's increasingly used by phishers to steal user's confidential information such as password, pin number etc., and perform business crime in recent years.

Within one to two years, the number of phishing attacks increased dramatically. Our analysis identifies that the phishing hyperlinks share one or more characteristics as listed below:

1) The visual link and the actual link are not the same;

2) The attackers often use dotted decimal IP address instead of DNS name;

3) Special tricks are used to encode the hyperlinks maliciously;

4) The attackers often use fake DNS names that are similar (but not identical) with the target Web site.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 5, March 2017

We then propose an end-host based anti-phishing algorithm which we call Link Guard, based on the characteristics of the phishing hyperlink. Since Link Guard is character-based, it can detect and prevent not only known phishing attacks but also unknown ones. We have implemented Link Guard in Windows XP, and our experiments indicate that Link Guard is light-weighted in that it consumes very little memory and CPU circles, and most importantly, it is very effective in detecting phishing attacks with minimal false negatives.

2.MAIN CHARACTERISTICS

Evolving with the anti-phishing techniques, various phishing techniques and more complicated and hard-to-detect methods are used by phishers. The most straightforward way for a phisher to defraud people is to make the phishing Web pages similar to their targets. Actually, there are many characteristics and factors that can distinguish the original legitimate website from the forged e-banking phishing website like Spelling errors, Long URL address and Abnormal DNS record. The full list is shown in table I which will be used later on our analysis and methodology study.

3. RELATED WORK

This section reviews the related works of phishing attacks. Various researches on phishing attack have been done for the past few years.

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. Other 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 incite 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. However, an automatic anti-phishing method is seldom reported. The typical technologies of anti-phishing from the User Interface aspect are done. 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. [8] discussed about a system, In this proposal, a neural network approach is proposed for energy conservation routing in a wireless sensor network. Our designed neural network system has been successfully applied to our scheme of energy conservation. Neural network is applied to predict Most Significant Node and selecting the Group Head amongst the association of sensor nodes in the network. After having a precise prediction about Most Significant Node, we would like to expand our approach in future to different WSN power management techniques and observe the results. In this proposal, we used arbitrary data for our experiment purpose; it is also expected to generate a real time data for the experiment in future and also by using adhoc networks the energy level of the node can be maximized. The

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 5, March 2017

selection of Group Head is proposed using neural network with feed forward learning method. And the neural network found able to select a node amongst competing nodes as Group Head.

4. PROPOSED SYSTEM

4.1 SYSTEM ARCHITECTURE

Our phishing detection system architecture is depicted below. There are four components in our system.

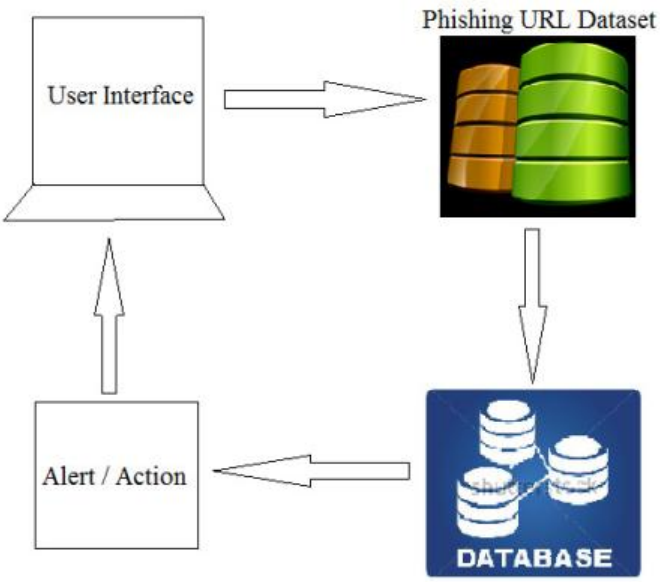

Figure 1 .System Architecture

- User Interface - The user interface (UI) is everything designed into an information device with which a person may interact
- Phishing URL Dataset - A dataset is a collection of data. Most commonly a data set corresponds to the contents of a single database table.
- Database - Database manages the Phishing URL dataset.
- Alert / Action - It is the output to the user's URL query.

4.2 DATA FLOW DIAGRAM

This diagram depicts the graphical representation of the "flow" of data through an information system, modelling its process aspects.

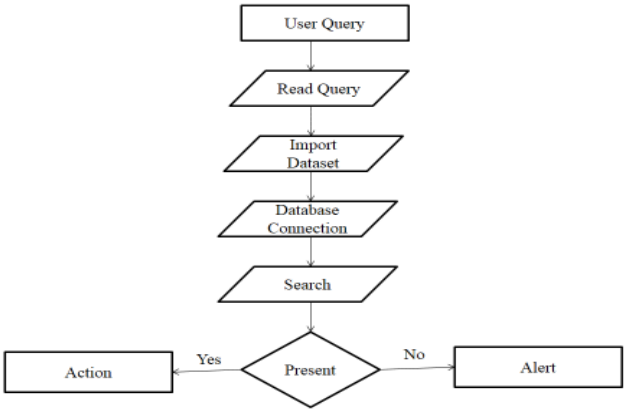

Figure 2 .Data Flow Diagram

1. The User's Query is the input to our system.
2. Input is collected to process it.
3. Phishing URL dataset is imported into the database
4. Database connection is made to perform database operations.
5. The features of the given URL such as IP address, URL length is validated.
6. If the features are matched with the features of Phishing Website then the user will get the alert.
7. Else the user's action will be performed.

4.3 MODULES OF PROPOSED SYSTEM

- User Interface
- Import dataset
- Database operations
- Alert/Action

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 5, March 2017

4.3.1 User Interface

The User Interface is the junction between a user and a computerprogram. An interface is a set of commands or menus through which a user communicates with a program. A command-driven interface is one in which you enter commands. A *menu-driven* interface is one in which you select command choices from various menus displayed on the screen. The user interface is one of the most important parts of any program because it determines how easily you can make the program do what you want. A powerful program with a poorly designed user interface has little value. Graphical user interfaces (GUIs) that use windows, icons, and pop-up menus have become standard on personal computers. Secured browser is developed as a user interface. User enters the URL to visit in our secured browser. User is allowed to visit the page only if the URL is valid.

4.3.2 Import Dataset

A data set is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question. The data set lists values for each of the variables, such as height and weight of an object, for each member of the data set. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows. The term data set may also be used more loosely, to refer to the data in a collection of closely related tables, corresponding to a particular experiment or event. An example of this type is the data sets collected by space agencies performing experiments with instruments aboard space probes. Our dataset consists of 2346 phishing URLs. The dataset is collects from Malware domain list website. They update the Phishing URLs in a regular interval. Main Syntax for import the dataset into the database

```
CALL SYSCS_UTIL.SYSCS_IMPORT_TABLE
(null,'SUGI','c: \export1.csv',';','%', null, 0);
```

4.3.3 Database Operations

A database is an organized collection of data. It is the collection of schemas, tables, queries, reports, views, and other objects. The data are typically organized to model aspects of reality in a way that supports processes requiring information, such as modelling the availability of rooms in hotels in a way that supports finding a hotel with vacancies. Database manages the Phishing URL dataset. Database connection is done to mine the attributes of dataset. Using database query, it checks whether the given URL is Phish or Legitimate website.

4.3.3. Action / Alert

An alert is a warning to people to be prepared to deal with something dangerous. Alert messaging (or alert notification) is machine-to-person communication that is important or time sensitive. An alert may be a calendar reminder or a notification of a new message.Alert / Action is the output to the user's URL query. The features of the given URL such as IP address, URL length is validated. If the features are matched with the features of Phishing Website then the user will get the alert. Else the user's action will be performed.

5. EXPERIMENTAL RESULTS

Figure shows the user interface and the dataset collection. It also depicts the alert and action that happen with respect to the user's input query.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 5, March 2017

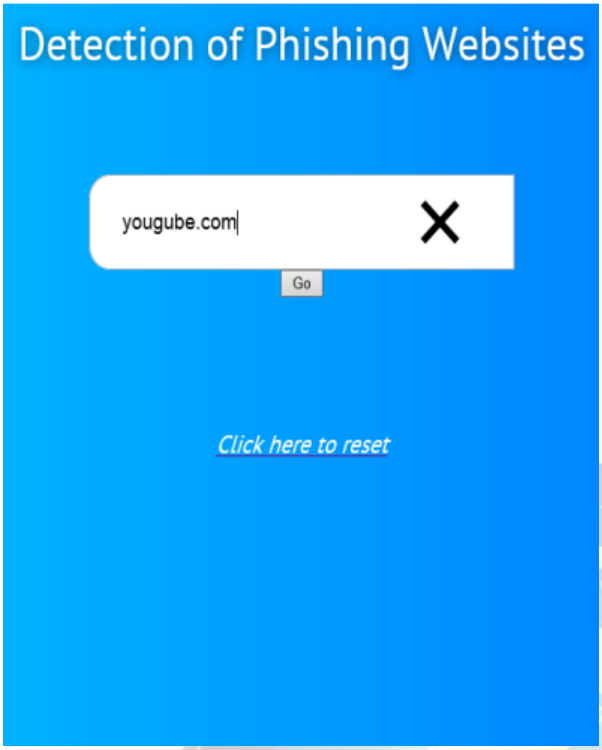

Figure 3. User Interface

Figure 4. Alert

Date (UTC)	Domain	IP	Reverse Lookup	Description	Registrant	ASN
2017/03/20_10:13	alegroup.info/ntnrrhst	194.87.217.87	mccfortwayne.org.	Ransom, Fake.PCN, Malspam	Lee Everton / lee_everton2002@yahoo.com	197695
2017/03/20_10:13	fourthgate.org/Yryzvt	104.200.67.194	-	Ransom, Fake.PCN, Malspam	Charlie Dillon / goddaddy@638united.com	8100
2017/03/20_10:13	dieutribenhkhop.com/parking/	84.200.4.125	125.0-255.4.200.84.in-addr.arpa.	Ransom, Fake.PCN, Malspam	-	31400
2017/03/20_10:13	dieutribenhkhop.com/parking/pay/rd.php?id=10	84.200.4.125	125.0-255.4.200.84.in-addr.arpa.	Ransom, Fake.PCN, Malspam	-	31400
2017/03/14_23:02	ssl-6582datamanager.de/	54.72.9.51	ec2-54-72-9-51.eu-west-1.compute.amazonaws.com.	redirects to Paypal phishing	goldanderwand@aol.com	16509
2017/03/14_23:02	privatkunden.datapipe9271.com/	104.31.75.147	-	Paypal phishing	Registrar Abuse Contact abuse@namecheap.com	13335
2017/03/06_21:09	www.hjaoopoa.top/admin.php?f=1.gif	52.207.234.89	ec2-52-207-234-89.compute-1.amazonaws.com.	Cerber ransomware	Registrant lecborbob i@rothtec.com	14618
2017/03/06_21:09	up.mykings.pw:8888/update.txt	60.250.76.52	60-250-76-52.HINET-IP.hinet.net.	related to a Mirai windows spreader trojan	Registrant 30da1310f05f42d7a349460c551aee6f.protect@whoisguard.com	3462
2017/03/06_21:09	down.mykings.pw:8888/ver.txt	60.250.76.52	60-250-76-52.HINET-IP.hinet.net.	related to a Mirai windows spreader trojan	Registrant 30da1310f05f42d7a349460c551aee6f.protect@whoisguard.com	3462
2017/03/06_21:09	down.mykings.pw:8888/ups.rar	60.250.76.52	60-250-76-52.HINET-IP.hinet.net.	related to a Mirai windows spreader trojan	Registrant 30da1310f05f42d7a349460c551aee6f.protect@whoisguard.com	3462
2017/02/09_14:04	fo5.a1-downloader.org/g2v9s1.php?id=yourname@yourdomain.com	188.225.32.177	vds-tibca.timeweb.ru.	trojan download	Protection of Private Person / a1-downloader.org@regprivate.ru	9123
2017/02/09_14:04	aadroid.net/sys.olk	107.180.51.15	ip-107-180-51-15.ip.secureserver.net.	ransomware	-	26496
2017/01/25_20:16	falconsafe.com.sg/api/get.php?id=aW5mb0BzYXBjdXBncmFkZXMuY29t	43.229.84.107	-	Trojan.Backdoor, Office.Word.Downloader	domain@exabytes.sg	38532
2017/01/25_20:15	www.lifelabs.vn/api/get.php?id=aW5mb0BzYXBjdXBncmFkZXMuY29t	118.69.196.199	-	Trojan.Backdoor, Office.Word.Downloader	-	18403
2017/01/19_13:05	6tkx.uk-insolvencydirect.com/sending_data/in_cgi/bbwp/cases/Inquiry.php	35.166.113.223	ec2-35-166-113-223.us-west-2.compute.amazonaws.com.	leads to ransomware	Registrar Abuse Contact abuse@namesilo.com	16509
	daralasnan.com/wp-co		sg2nlhg800c1800.shr.			

Figure 5. Dataset Collection

6. CONCLUSION AND FUTURE WORK

Phishing has becoming a serious network security problem, causing finical loss of billions of dollars to both consumers and e-commerce companies. And perhaps more fundamentally, phishing has made e-commerce distrusted and less attractive to normal consumers. In this paper, we have studied the characteristics of the hyperlinks that were embedded in phishing e-mails.

We then designed an anti-phishing algorithm, Link-Guard, based on the derived characteristics. Since Link-Guard is characteristic based, it can not only detect known attacks, but also is effective to the unknown ones. We have implemented Link Guard for Windows XP. Our experiment showed that Link Guard is light-weighted and can detect up to 96% unknown phishing attacks in real-time. We believe that Link Guard is not only useful for detecting phishing attacks, but also can shield users from malicious or unsolicited links in Web pages and Instant messages.

As we have implemented this approach by considering the URL and Domain Identity Criteria, there are the different criteria needs to work in future.

References

[1] Phoebe Barraclough and Graham Sexton,"Phishing Website Detection Fuzzy System Modelling",Science and Information Conference ,July 2015.

[2] B. S. Osareh, "Intrusion Detection in Computer Networks based on Machine Learning Algorithms", IJCSNS International Journal of Computer Science and Network Security, vol. 8, November 2008.

[3] Abdulghani Ali ahmed,andNurulAmirahAbdullah,"Real Time Detection of Phishing Websites",IEEE conference ,2016.

[4] InsoonJo,Eunjin (EJ) Jung and Heon Y. Yeom,"You're not who you claim to be: website identity check for phishing detection",IEEE conference ,2010.

[5] Choon Lin Tan, Kang LengChiew and San Nah Sze, "Phishing Website Detection Using URL-Assisted Brand Name Weighting System"IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS) December 1-4, 2014.

[6] Anti-Phishing Working Group, Global Phishing Survey: Trends and Domain Name Use in 1H2009, http://www.antiphishing.org/reports/APWG globalPhishingSurvey 1H2009.pdf, October 2009.

[7] AninditaKhade and Dr. Subhash K Shinde,"Detection of Phishing Websites Using Data Mining Techniques",International Journal of Engineering Research & Technology (IJERT,Vol. 2 Issue 12, December – 2013.

[8] Christo Ananth, A.Nasrin Banu, M.Manju, S.Nilofer, S.Mageshwari, A.Peratchi Selvi, "Efficient Energy Management Routing in WSN", International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE), Volume 1, Issue 1, August 2015,pp:16-19

[9] Zhang, Y., Hong, J. I., &Cranor, L. F. (2007, May), "content-based approach to detecting phishing websites",In Proceedings of the 16th international conference on World Wide Web (pp. 639-648). ACM.

[10] Wikipedia, Phishing, http://en.wikipedia.org/wiki/Phishing.

[11] Y. Zhang, S. Egelman, L. Cranor, and J. Hong, Phinding phish: Evaluating anti-phishing tools, In the 14th Annual Network and Distributed System Security Symposium (NDSS 2007), 2007.

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