International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Research Paper
Available online at: www.jrrset.com
UGC Approved Journal No: 45483
Received on: 12.11.2017 Published on: 30.11.2017 Pages : 30-39

ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
Volume 5, Issue 11
Novmber 2017.
JIR IF : 2.54
DIIF IF : 1.46
SJIF IF : 1.329

PHISHING WEBSITE DETECTION USING ASSOCIATIVE CLASSIFIERS

Divya V[1], Vivitha Vijay[2]
[1,2] Assistant Professor, Department of Computer Science and Engineering, Sri Vellappally Natesan College of Engineering, Kerala, India

Abstract

Phishing is a criminal technique employing both social engineering and technical subterfuge to steal consumer's personal identity data and financial account credential. The aim of the phishing website is to steal the victim's personal information by visiting and surfing a fake webpage that looks like a true one of a legitimate bank or company and asks the victim to enter personal information such as their username, account number, password, credit card number, etc. The main goal is to investigate the potential use of automated data mining techniques in detecting the complex problem of phishing websites in order to help all users from being deceived or hacked by stealing their personal information and passwords leading to catastrophic consequences. Experimentations against phishing data sets and using different common associative classification algorithms (MCAR and CBA) and traditional learning approaches have been conducted with reference to classification accuracy.

Keywords: Phishing Websites, Data Mining, Associative Classification, Machine Learning.

1. Introduction

During the last decade, most of the financial and government organizations have extended their online services to their clients. In 2011, 83% of Americans and 85% of Europeans regularly shopped online. With the emerging use of smart phones, increasing number of people are depending on online services to shop, check their banking account, pay their bills, or even play with anonymous friends. While such activities had an important impact on the world economy, such large dependencies on online financial services increases security risks for both customers and financial institutes.

Phishing is a criminal technique employing both social engineering and technical subterfuge to steal consumer's personal identity data and financial account credential. Phishing is a new identity theft crime. The media reports stories almost on a daily basis about an organization that has customers targeted by a phishing attack. While financial organizations try always to improve their security techniques in order to protect their customers, phishers develop even more sophisticated attacking techniques.

Phishing websites [9] are fake web pages that are created by malicious people to imitate web pages of real websites. Phisher typically create web pages that are visually very similar to the real web pages in order to scam their victims. An unaware client might be easily deceived by this kind of scam. The Victims of a phishing Web page may expose their bank account, password, credit card number, or other important information to the phishing Web page owners. While phishing is a relatively new Internet crime when compared to other forms (e.g., viruses and hacking), a recognizable increase in the number and severity of phishing attacks is reported. According to a recent study by Gartner (2011), 57 million US Internet users have identified the receipt of email linked to phishing, about 1.7 million of them are thought to have yielded to the convincing attacks and tricked them into revealing personal information. Studies by the Anti-Phishing Working Group (APWG) have concluded that Phishers are likely to succeed with as much as 5% of all message recipients.

The aim of the phishing website is to steal the victims' personal information by visiting and surfing a fake webpage that looks like a true one of a legitimate bank or company and asks the victim to enter personal information such as their username, account number, password, credit card number, …,etc. The impact is the break of information security through the compromise of confidential data and the victims may finally suffer losses of money or other kinds of assets. The attackers might also commit identity theft crimes using the victim's stolen information. Moreover, phishing attacks also damage the reputation of the attacked financial institutes since customers become less confident that they can securely access their accounts. Therefore, they might switch to other institutes.

Phishing has a huge negative impact on organizations' revenues, customer relationships, marketing efforts, and overall corporate image. Phishing attacks may cost companies hundreds of thousands of dollars per attack in fraud-related losses and personnel time. Even worse, costs associated with the damage to brand image and consumer confidence can run in the millions of dollars.

Due to the wide variety of data being captured, efficient management and quick retrieval of information is very important for decision making. Data mining is the science of extracting meaningful information from these large data sets (Witten and Frank, 2000). Data mining and knowledge discovery techniques have been applied to several areas including market analysis, industrial retail, decision support and financial analysis.

This paper's main goal is to investigate the potential use of automated data mining techniques in detecting the complex problem of phishing Websites. This type of prediction is closely related to classification problem in data mining where the class attribute in this case is the degree of phishing. The classification process will be based on the different characteristics such as spelling errors, long URLs, personalization, prefix and suffix, etc. collected from the input Websites using different online tools.

2. Literature survey

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 website phishing 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 proposed is the integration of classification rule mining and

association rule mining algorithm, the main goal of this algorithm is to categorize the key factors in detecting phishing website. Most of the phishing attacks use emails in order to fake their victims. According to this approach, the phishing problem can be considered as a spam filtering problem and therefore can be handled with effective spam filters.

There are several promising defending approaches to this problem reported earlier. 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 IE8 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. Sensitive information that is not related to a specific site, e.g., credit card information and SSN (Social Security Number), 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. Authors in 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.

Another approach is to employ certification, e.g., Microsoft spam privacy. A recent and particularly promising solution was proposed in, which combines the technique of standard certificates with a visual indication of correct certification. 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 Net craft, Web sense and Cloud mark anti-phishing toolbars, prevents phishing attacks by utilizing a centralized blacklist of current phishing URLs. The weaknesses of this approach are its poor scalability and its timeliness. The typical technologies of anti-phishing from the user interface aspect are done by. 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. In the visual similarity of Web pages is oriented, and the concept of visual approach to phishing detection was first introduced.

3. Existing system

During the last decade, most of the financial and government organizations have extended their online services to their clients. In 2011, 83% of Americans and 85% of Europeans regularly shopped online. With the emerging use of smart phones, increasing number of people are depending on online services to shop, check their banking account, pay their bills, or even play with anonymous friends. While such activities had an important impact on the world economy, such large dependencies on online financial services increases security risks for both customers and financial institutes.

The aim of the phishing website is to steal the victims' personal information by visiting and surfing a fake webpage that looks like a true one of a legitimate bank or company and asks the victim

to enter personal information such as their username, account number, password, credit card number, …,etc. The impact is the break of information security through the compromise of confidential data and the victims may finally suffer losses of money or other kinds of assets. The attackers might also commit identity theft crimes using the victim's stolen information. Moreover, phishing attacks also damage the reputation of the attacked financial institutes since customers become less confident that they can securely access their accounts. Therefore, they might switch to other institutes.

4. Proposed description

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 website phishing attacks, there are several promising approaches to this problem and a comprehensive collection of related works. This paper main goal is to investigate the potential use of automated data mining techniques in detecting the complex problem of phishing Websites. This type of prediction is closely related to classification problem in data mining where the class attribute in this case is the degree of phishing. The classification process will be based on the different characteristics such as spelling errors, long URLs, personalization, prefix and suffix, etc. collected from the input Websites using different online tools. The motivation behind this project is to create a resilient and effective method that uses data mining algorithms and tools to detect phishing websites in the Artificial Intelligent technique. Association and classification algorithms can be very useful in predicting phishing websites.

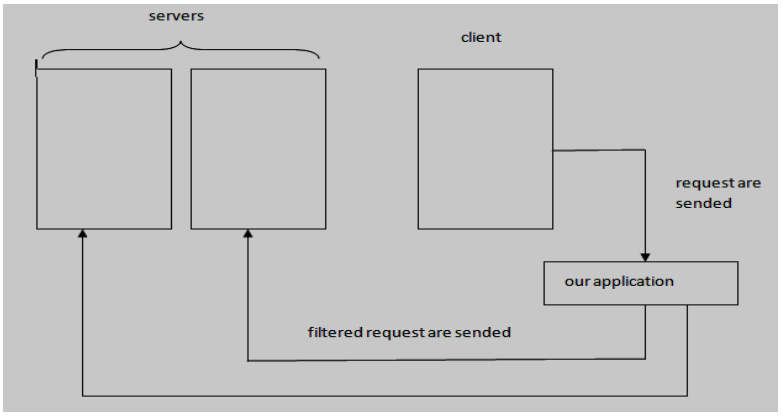

Fig 4.1: Project in one figure

Phishing websites is a semantic attack which targets the user rather than the computer. It is a relatively new Internet crime in comparison with other forms, e.g., virus and hacking. The phishing problem is a hard problem because of the fact that it is very easy for an attacker to create an exact replica of a good banking site, which looks very convincing to users. The word phishing from the phrase "website phishing" is a variation on the word "fishing". The idea is that bait is thrown out with the hopes that a user will grab it and bite into it just like the fish. In most cases, bait is either an e-mail or an instant messaging site, which will take the user to hostile phishing websites. The motivation behind this study is to create a resilient and effective method that uses Data Mining algorithms and tools to detect e-banking phishing websites in an Artificial Intelligent technique. Associative and classification algorithms can be very useful in predicting Phishing websites. It can give us answers about what are the most important e-banking phishing website characteristics and indicators and how they relate with each other. Comparing between different Data Mining classification and association methods and

techniques is also a goal of this investigation since there are only few studies that compares different data mining techniques in predicting phishing websites.

Phishing is the attempt to obtain sensitive information such as usernames, passwords, and credit card details, often for malicious reasons, by disguising as a trustworthy entity in an electronic communication. The word is a neologism created as a homophone of fishing due to the similarity of using a bait in an attempt to catch a victim. According to the 3rd Microsoft Computing Safer Index Report released in February 2014, the annual worldwide impact of phishing could be high as $5 billion. Phishing is typically carried out by email spoofing or instant messaging, and it often directs users to enter personal information at a fake website, the look and feel of which are almost identical to the legitimate one. Communications purporting to be from social web sites, auction sites, banks, online payment processors or IT administrators are often used to lure victims. Phishing emails may contain links to websites that are infected with malware.

Phishing is an example of social engineering techniques used to deceive users, and exploits weaknesses in current web security. Attempts to deal with the growing number of reported phishing incidents include legislation, user training, public awareness, and technical security measures. Many websites have now created secondary tools for applications, like maps for games, but they should be clearly marked as to who wrote them, and users should not use the same passwords anywhere on the internet.

5. Phishing types

5.1. Spear phishing

Phishing attempts directed at specific individuals or companies have been termed spear phishing. Attackers may gather personal information about their target to increase their probability of success. This technique is, by far, the most successful on the internet today, accounting for 91% of attacks.

5.2. Clone phishing

Clone phishing is a type of phishing attack whereby a legitimate, and previously delivered, email containing an attachment or link has had its content and recipient address taken and used to create an almost identical or cloned email. The attachment or link within the email is replaced with a malicious version and then sent from an email address spoofed to appear to come from the original sender. It may claim to be a resend of the original or an updated version to the original. This technique could be used to pivot from a previously infected machine and gain a foothold on another machine, by exploiting the social trust associated with the inferred connection due to both parties receiving the original email.

5.3. Whaling

Several phishing attacks have been directed specifically at senior executives and other high-profile targets within businesses, and the term whaling has been coined for these kinds of attacks. In the case of whaling, the masquerading web page/email will take a more serious executive-level form. The content will be crafted to target an upper manager and the person's role in the company. The

content of a whaling attack email is often written as a legal subpoena, customer complaint, or executive issue. Whaling scam emails are designed to masquerade as a critical business email, sent from a legitimate business email, sent from a legitimate business authority. The content is meant to be tailored for upper management, and usually involves some kind of falsified company-wide concern. Whaling phisher men have also forged official-looking FBI subpoena emails, and claimed that the manager needs to click a link and install social software to view the subpoena.

5.4. Filter evasion

Phishers have even started using images instead of text to make it harder for anti-phishing filters to detect text commonly used in phishing emails. However, this has led to the evolution of more sophisticated anti-phishing filters that are able to recover hidden text in images. These filters use OCR, (optical character recognition) to optically scan the image and filter. Some anti-phishing filters have been used IWR (intelligent word recognition), which is not meant to completely replace OCR, but these filters can even detect cursive, hand-written, rotated (including upside-down text), or distorted text, as well as text on colored backgrounds.

5.5. Website forgery

Once a victim visits the phishing website, the deception is not over. Some phishing scams use Java Script commands in order to alter the address. An attacker can even use flaws in a trusted website's own scripts against the victim. These types of attacks are particularly problematic, because they direct the user to sign in at their bank or service's own web page, where everything from the web address to the security certificates appears correct. In reality, the link to the website is crafted to carry out the attack, making it very difficult to spot without specialist knowledge.

A Universal Man-in-the-middle phishing kit provides a simple-to-use interface that allows a phisher to convincingly reproduce websites and capture and capture log-in details entered at the fake site. To avoid anti-phishing techniques that scan websites for phishing-related text, phishers have begun to use Flash-based websites. These look much like the real website, but hide the text in a multimedia object.

5.6. Covert redirect

Covert redirect is a subtle method to perform phishing attacks that makes links appear legitimate, but actually redirect a victim to an attacker's website. Normal phishing attempts can be easy to spot because the malicious page's URL will usually be different from the real site link. For covert redirect, an attacker could use a real website instead by corrupting the site with a malicious login popup dialogue box.

5.7. Phone phishing

Not all phishing attacks require a fake website. Messages that claimed to be from a bank told users to dial a phone number regarding problems with their bank accounts. Once the phone number was dialed, prompts told users to enter their accounts numbers and PIN. Phishing sometimes uses fake caller-ID data to give the appearance that calls come from a trusted organization. SMS phishing uses cell phone text messages to induce people to divulge their personal information.

Other techniques

- Another attack used successfully is to forward the client to a bank's legitimate website, then to place a popup window requesting credentials on top of the page in a way that makes many users think the bank is requesting this sensitive information.

- Tab nabbing takes advantage of tabbed browsing, with multiple open tabs. This method silently redirects the user to the affected site. This technique operates in reverse to most phishing techniques in that it doesn't directly take the user to the fraudulent site, but instead loads the fake page in one of the browser's open tabs.

- Evil twin is a phishing technique that is hard to detect. A phisher creates a fake wireless network that looks similar to a legitimate public network that may be found in public places such as airports, hotels or coffee shops. Whenever someone logs on to the bogus network, fraudsters try to capture their passwords and/or credit card information.

6. Phishing website methodology

6.1. Data mining techniques

We utilized data mining classification and association rule approaches in our new phishing website detection model to find the most important phishing features and significant patterns of phishing characteristic or factors in the phishing website archive data. Particularly, we used a number of different existing data mining association and classification techniques including JRip [2], PART[2], PRISM[3] and C4.5, CBA [7], MCAR[8] algorithms to learn and to compare the relationships of the different phishing classification features and rules. The experiments of C4.5, RIPPER, PART and PRISM algorithms were conducted using the WEKA software system, which is an open java source code for the data mining community that includes implementations of different methods for several different data mining tasks such as classification, association rule and regression. CBA and MCAR experiments were conducted using an implementation version provided by the authors of .We have chosen these algorithms based on the different strategies they use to generate the rules and since their learnt classifiers are easily understood by human.

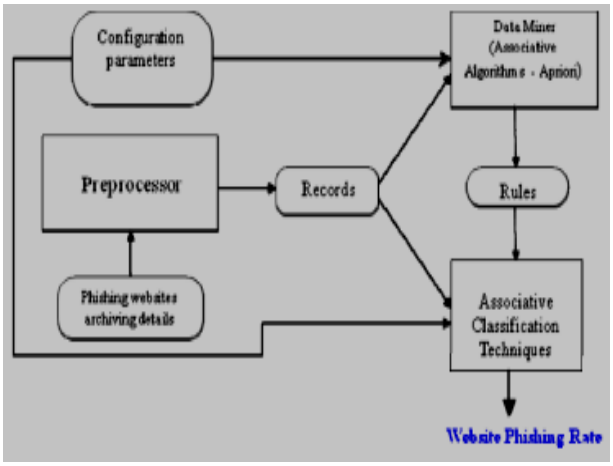

Fig 6.1: AC Model for detecting phishing websites

We used two web access archives, one from APWG [1] archive and one from Phish tank archive[5]. We managed to extract the whole 27 phishing security features and indicators and clustered them to its 6 corresponding criteria such as URL & Domain Identity, Security &Encryption, Source code &Java Script, Page Style & Contents, Web Address Bar and Social Human Factor.

6.2. Website phishing training data sets

Two publicly available datasets were used to test our implementation: the "phish tank" from the phishtank.com[4] which is considered one of the primary phishing report collates both the 2007 and 2008 collections. The Phish Tank database records the URL for the suspected website that has been reported, the time of that report, and sometimes further detail such as the screenshots of the website, and is publicly available. The Anti-Phishing Working Group (APWG) which maintains a "Phishing Archive" describing phishing attacks dating back to September 2007. A data set of 1006 phishing, suspicious and legitimate e-banking websites is used. In addition, 27 features are used to train and test the classifiers. We used a series of short scripts to programmatically extract the above features, and store these in an excel sheet for quick reference. Our goal is to gather information about classifying and categorizing of all different e-banking phishing attacks techniques. By thoroughly investigating these phishing attacks we've created a data set containing information regarding what different techniques have been used and how it can be predicted.

6.3. Mining e-banking phishing challenges

The age of the dataset is the most significant problem, which is particularly relevant with the phishing corpus. E-banking Phishing websites are short-lived, often lasting only in the order of 48 hours. Some of our features can therefore not be extracted from older websites, making our tests difficult. The average phishing site stays live for approximately 2.25 days[5]. Furthermore, the process of transforming the original phishing website archives into record feature row data sets is not without error. It requires the use of heuristics at several steps[6].

7. Implementation

The system consists of four modules:

7.1 Web portal scanner

In web portal scanner, it will check all incoming and outgoing packets to find out phishing website access or attacks. With the help of an automated ARP resolution handler we have option to identify the web url/ip address using associative classification. We found out that the MCAR algorithm scales well if compared to common classification data mining algorithms. In particular, MCAR has achieved on average 6.8%, 6.1% and 5.4% higher accuracy than SVM.CBA and NB respectively. And CBA algorithm outperformed SVM and NB algorithms. Associative classifiers approach produces more accurate classification accuracy than other traditional classification approaches such as statistical, probabilistic.

Also the rules generated from our associative classifier(MCAR) indicates that URL and domain identity, and security and encryption features are consider important features to increase the

final detection rate. The experiments demonstrate the feasibility of using associative classification techniques in real applications involving large datasets.

7.2 Internal database handler

Internal database handler is using to update phishing websites detection according to URL and IP address or meta data information corresponding to portal. Each and every time it will synchronize between apps installed in each users system.

7.3 Packet scanner

Packets scanner is using to identify type of packets and behavior. we have also option to identify and separate protocols according to each packets scanner move the operation process to network analyzer to find out whether it is spam ware or not. Packet scanner is also w\using to set upload limit/protocols.

7.4 Network analyzer

The network analyzer will scan all the incoming and outgoing request on client and it will work according to types and rules related to phishing website detector with the help of data mining and associative classification. The main goal is to stop the personal information breakings through wide area network and unauthorized system folder access across ftp over law or via add-ons.

8. Conclusion

Phishing is a criminal technique employing both social engineering and technical subterfuge to steal consumer's personal identity data and financial account credential. Phishing is a new identity theft crime. Phishing website model based on classification data mining showed the significance importance of the phishing website two criteria's (URLL & domain identity) and (security & encryption) in the final phishing detection rate, and also showed the insignificant trivial influence of some other criteria like ' page style &content' and 'social human factor' in the final phishing rate. The rules generated from the associative classification model showed the correlation and relationship between some of their characteristics which can help us in building phishing website detection system. The experiments demonstrate the feasibility of using associative classification techniques in real applications involving large databases and its better performance as compared to other traditional classification algorithms. MCAR performs better than other algorithm in terms of accuracy. Experimentation against phishing data sets using different classification algorithms have been performed. The base is the accuracy measure. The results obtained reveal that the MCAR algorithm outperformed all other algorithms with respect to accuracy. As for future work, we want to use different pruning methods like lazy pruning which discards rules that incorrectly classify training instances and keeps all other rules to be used by MCAR associative classification technique in order to minimize the size of the resulting classifiers and to experimentally measure and compare the effect of these different pruning on the final result.

References

[1] Anti-Phishing Working Group. Phishing Activity Trends Report, http://antiphishing.org/apwg_report_final.pdf. 2007.
[2] I.H. Witten and E. Frank, "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, CA, 2005.
[3] J. Cendrowska., "PRISM: An algorithm for inducing modular rule", International Journal of Man-Machine Studies (1987), Vol.27, No.4, pp.349-370.
[4] FDIC, Tech. Rep., "Putting an end to account-hijacking identity theft", Dec.2004.[Online]. Available:http://www.fdic.gov/consumers/idtheftstudy/identity theft.pdf.
[5] Ian Fette, Norman Sadeh and Anthony Tomasic, "Learning to Detect Phishing Emails", Institute for Software Research International, CMU-ISRI-06-112, June 2006.
[6] Bing Liu, Wynne Hsu, Yiming Ma, "Integrating Classification and Association Rule Mining." Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98, Plenary Presentation), New York, USA, 1998.
[7] F.Thabtah, P. Cowling and Y. Peng, "A new multi-class,multi-label associative classification approach".The 4th International Conference on Data Mining(ICDM'04), Brighton, UK, 2004.
[8] L. James, "Phishing Exposed," Tech Target Article by: Sunbelt software, searchexchange.com, 2006.