Model for Detection of Phishing Sites using Clustering and Bayesian Approach: A Survey

Nilima Ramdas Narad
D.Y.Patil College Of Engineering-DYPCOE
Computer Engineering, Ambi, Pune, India

Sandeep U Kadam
D.Y.Patil College Of Engineering-DYPCOE
Computer Engineering, Ambi, Pune, India

ABSTRACT

Web Phishing is a major attack nowadays.web phishing is phisher tries to get users sensitive information like bank details, ATM pin or any personal information. After extracting users information attacker may misuse this information. Nowadays most of the people prefer online shopping or online payment and user has to share his personal or sensitive information on web page. User may not think about website security. So website security is very important. Before doing any transaction and sharing any personal information on web page, user must ensure the security of that website. The best solution for this problem is to protect from phishing is to identify a phish. Phishing emails usually seem to come from well-known organization and ask your personal information such as credit card number, security number, account number or passwords. What actually attacker does? The attacker creates the no of replicas of authenticate sites, and users are forced to direct to that websites by attracting them with offers. As standard mentioned in W3C (World Wide Web Consortium), I am proposing a system which can easily recognize the difference between authenticate site and phishing site. There are certain standards which are given by W3C (World Wide Web Consortium), based on these standards I am choosing some features which can easily describe the difference between legit site and phish site. To protect you from phishing, I am proposing a model to determine the fraud sites. To determine the phishing attack, URL features and HTML features of web page are considered. Clustering algorithm such as K-Means clustering is applied on the database and prediction techniques such as Naive Bayes Classifier is applied. By applying this, probability of the web site as valid Phish or Invalid Phish. To check the validity of URL, if still user is not able decide the validity of web page then Naïve Bayes Classifier is applied.

Keywords

Anti Phishing Technique, Bayesian Approach, Data Mining, Database Clustering, and Phishing Attacks.

1. INTRODUCTION

Phisher is the community of attacker which creates the replicas of the legitimate web sites to submitting user’s personal information such as passwords, credit card number, and financial transaction information to illegitimate websites [1]. Since the last December 2012 to January 2013, there is rise in phishing attacks by 2% as described in survey of RSA fraud Surveyor [2]. The W3C has set some standards, specifications and recommendations that are followed by most of the authenticate sites. But a phisher may not care to follow these standards as this site is intended to catch many fish in very small amount of time and bait [6]. For prevention and detection of attacks various preventive strategies are developed by most common anti-phishing service provider such as Google Toolbar, an antivirus provider [3].What actually this service provider does? This service provider creates and maintains the database of sites which are blacklisted. There are some organizations like http://www.phishtank.com/ which are anti-phishing organizations. These organizations keep the record of blacklisted sites or phishing sites. There are various techniques are available for detection of phish, such as, plug-In-browser .This techniques maintains the online repositories of blacklisted sites. The phisher always creates the site at such a rate that in a particular time period that site is not reported as phish, in that case these techniques fails. By observatation, the major disadvantages of is like the normal user will not always take the precaution of phishing site. Due to the overall look of site like legitimate site and this may happen this site is not blocked by service provider.

I am proposing his system to escape from phishing and to overcome the disadvantages of existing systems. I have proposed an efficient method to detect the phishing sites. My model differentiates the phishing site and authenticates sites. Model uses the URL features [3] and HTML features. To check the validity of the site, K-Means Clustering and Naive Bayes Classifier [4] used. The K-Means Clustering is applied on the URL features of the web site and the feature set is plotted in one of the two clusters of database. If the feature set is nearest to more suspicious then site is declared as Phishing, if site is nearest to less suspicious then it is a authenticate site but if the feature set not nearer to less suspicious cluster or more suspicious cluster it means it is nearer to cluster in between them. If the site is in the nearer to the cluster between them then there is need of more feature extraction where you will extract HTML features by using DOM representation [5] of the HTML and features of different tags are observed. A Naive Bayes Classifier is employed if K-Means clustering is not that much useful, considering both URL and HTML features and the training datasets provided to predict the legit site or phish site.

2. URL FEATURES AND HTML FEATURES

2.1 URL Features

For detection of phishing sites consider the features of URL. As per W3C standard, following are the URL features:

- **Dots:** More number of DOTS in URL, more chances of site being phish. Phisher uses the fake domain to create the legitimate look of URL by using more no of dots in URL.

- **IP Address:** The domain name is one of the pieces inside the URL, if is an entire set of directions and it contains extremely detailed information. It is also the most easily recognized part of the entire address. IP address needs to be registered. Authenticate site have their registered domain. This is an important feature to recognize phishing sites.

- **Suspicious Characters:** Phisher can use the extra features of URL to fool the user. He can use some special character like '&', '-', '@','_'.by using this special character hacker creates the look like a legitimate site and automatically user easily click on.
- **Slashes:** Existence of sub-folders in URL is the presence of slashes in the URL.The purpose of this subfolder is to hide the information.

2.2 HTML features

Attacker copies the source code of legitimate site to his own page. And then tries to modify the page so that it becomes more similar to legitimate site. Only URL features cannot predict whether site is phishing or not. When it's not possible to predict phishing sites then need to extract more features of site called as HTML features. HTML features are extracted from source code. For extraction of HTML features you require HTML DOM-tree parser [8].

following are the HTML features:

- SSL Certificate: It is Secure Socket Layer Certificate issued form some authorized organizations like W3C.it gives the unique identity of owner of web site with detail information of how it is encrypted. Every authenticate site have SSL certificate version of 2.0 and 3.0. Validity of SSL certificate needs to be updates as it has validity of very short period. And needs to be updated over period of time. Without SSL certificate most of the browsers will deny page access. Phisher have very less chances of getting SSL certificate because it's given to only authenticate sites.
- Foreign tags: After clicking on web page this web page redirect to another domain. This domain will not belong to any domain or sub-domain of current site. Many websites generally have same foreign links but when there is more number of foreign links, this will increase the unsavory about that site.
- NULL tags: A NULL tag doesn't return anything. After clicking on such link, but it no longer shows up as a link, it returns the same page. When clicked on such links nothing happens or the links are redirected to the same page. Phisher copies the source code of legitimate; he may remove the most of the links. Presence of More the NULL tags, more chances of site being phish.

3. LITERATURE SURVEY

Current Phishing Status

Looking at the First fortnight report by Anti-Phishing Organization (www.antiphishing.org) and RSA Online Fraud Attacks Surveys few major points:

- Phishing attacks have been increased by 2% since December 2012.
- India is having 4% of global attacks by volume of attack.
- India is being targeted 4% of global attacks by volume of brands attacked.

Taking the reference of phishing activity trends report, 2nd quarter (2014) produced by APWG (antiphishing work group) few major points are noticed:

- Total 128,378 sites were observed as phishing sites.
- Most targeted and more frequently industries are online payment and crypto-surrency.
- Increase in PUPs (Potentially unwanted programs) and this leads to higher global infection
- One of the top hosting phishing site country is United states.

Fig.1 shows the most targeted industry with their total percentage. Few major points observed in above diagram:

- As like with the previous phishing report payment services is the most targeted industry with 39.80% of attacks.
- 2nd targeted industry is financial industry with 20.20% of attacks.
- 3rd targeted industry in Retail/services industry with percentage 16.53 of attacks.

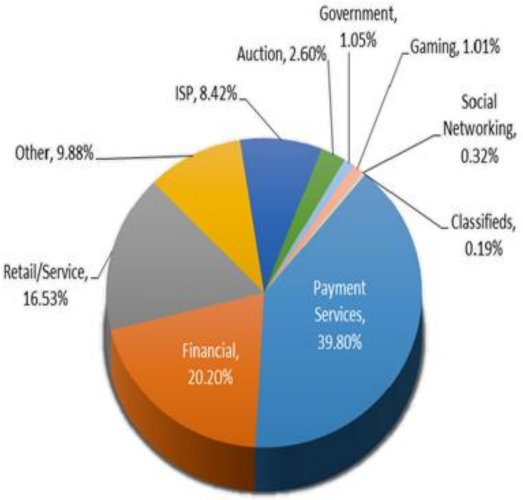

Fig.1: Most targeted industry sector

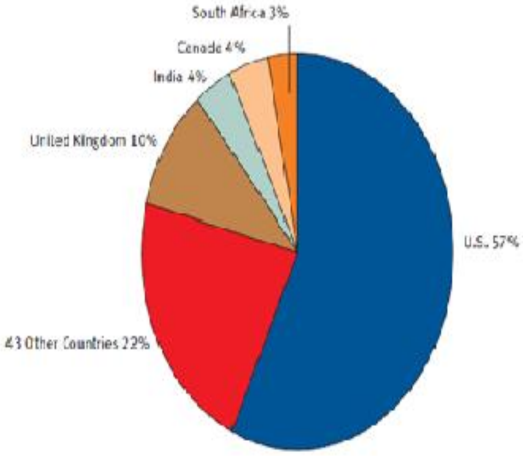

Fig.2: Major attacked countries by volume of attack

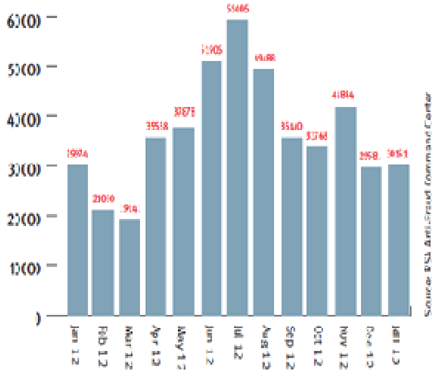

Fig.3: Total reported attacks per month for 1 year

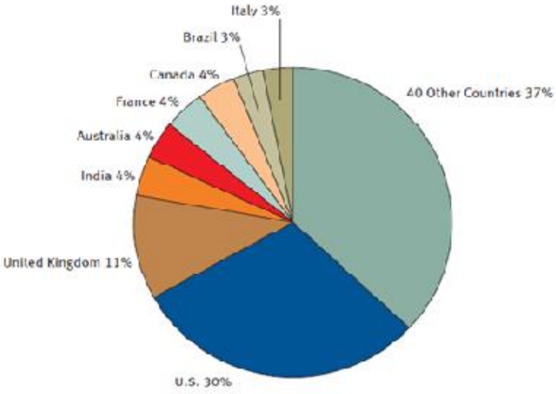

Fig.4: Major attacked countries by Brands attacks

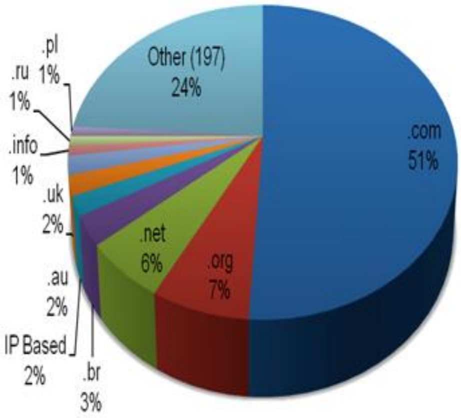

Fig.5: Phishing By TLD, second quarter 2011

4. CONCLUSION AND FUTURE SCOPE

Phishing attacks are increasingly rapidly. There is need to develop techniques for detection and prevention of phishing sites. In this survey paper, there is phishing detection a system mechanism out of which one is dependent on URL features of web-sites and second is based on HTML tags and Visual Features of web-sites. Technique is based on system which is a trail of combination of these two mechanisms and using base techniques given by them. Application of clustering on this system generates the output faster but by compromising with the accuracy of results. Bayesian approach generates more accurate results but it requires analyzing the training data set provided and takes a very long time of execution. Systems have used a combination of these two algorithms resulting into a mechanism which is more efficient and reliable than these two separate techniques. Mechanism uses K-Means Clustering which is efficient to generate output at higher throughput but with lack of efficiency and this lack of efficiency is recovered with the Naive Bayes Classifier. Step by step instructions to develop our schema for these extra issues is under our examination. The given paper is more concerned with NP-hard problems; more studies are needed for the problems which are not NP-Hard.

5. ACKNOWLEDGMENT

I am deeply indebted to my project guide Prof. Sandeep Kadam for his constant help, stimulating suggestions and encouragement. I would like to express my gratitude to all those who gave me the possibility to complete this survey paper.

6. REFERENCES

[1] RachnaDhamija, J. D. Tygart, and Marti Heast, “Why Phishing Works”, CHI-2006, Conference on Human Factor in Computing Systems, April 2006.

[2] Phishing Activity Trend report 2nd quarter 2014,http://www.apwg.org.

[3] RSA Online Fraud Surveyor, “The phishing kit – the same wolf, just different sheep’s clothing”, RSA Surveys, vol-1, February-2013.

[4] Xiaoping GU, Hong Yuan WANG, and Tonguing NI “An Efficient Approach to Detect Phishing Web” Journal of Computational Information Systems 9:14(2013), 2013, pp. 5553-5560.

[5] Computational Information Systems 9:14(2013), 2013, pp. 5553-5560.

[6] Haijun Zhang, Gang Liu, Tommy W. S. Chow, Senior Member, IEEE, and Wenyin Liu, Senior Member, IEEE “Textual and Visual Content-Based Anti-Phishing: A Bayesian Approach”, vol-22, IEEE Transactions October- 2011 pp. 1532-1546.

[7] Angelo P. E.Rosiello, EnginKirda, Christopher Kruegel, FabrizioFerrandi, and Politecnico di Milano “A Layout-Similarity-Based Approach for Detecting Phishing Pages”- unpublished

[8] WIKIPEDIA.ORG- The Online Encyclopedia, http://www.wikipedia.org/

[9] Abraham Sillberschatz, Henry Korth, and S. Sudarshan, “Database System Concepts”, 5th Edition, pp. 900-903.

[10] PHISHTANK.COM- The Online Valid Phish Sites Repository,http://data.phishtank.com/data/online-valid.csv