Detection of Phishing Websites using URL and terms in the Webpage

Chris D’Souza
PG student: Computer Dept.
Shree L. R. Tiwari College of Engg.
Thane, India
chrisroopal@gmail.com

Vinayak Shinde
HOD & Professor: Computer Dept.
Shree L. R. Twari College of Engg.
Thane, India
vdshinde@gmail.com

Abstract**—** Phishing is the combination of social engineering and technical exploits used to convince a victim to provide their personal information, usually for financial gains [10]. Phishing has become the most popular practice among cyber criminals due to which they are becoming more frequent and more sophisticated. The impact of phishing is drastic and significant since it involves the risk of identity theft and moreover financial losses. Phishing scams have become a bane for net-banking and e-commerce users. Many anti-phishing mechanisms currently focused to verify whether a web site is genuine or not. In this paper, study of anti-phishing techniques to safeguard users against them is elaborated.

Keywords**—** *Brand name; Phishing E-mail; phishing detector; hyperlinks; Search Engine; Term Frequency-Inverse Document Frequency.*

*****

I. Introduction

The Internet plays an increasingly significant role in today’s commerce and business activities. Unfortunately, poor security on the Internet and the opportunity for large financial gains provide a strong motivation for attackers to perpetrate seemingly low risk, yet high-return online scams. Email messages are less protected as they move across the Internet. Often information being transmitted is valuable and sensitive thus necessitating effective protection mechanisms to prevent information from being abused or to protect confidential information from being revealed to unauthorized parties.

The phishing webpage is a replica of other sites that look like legitimate one. Phishing is a heinous crime like hacking and is also known as brand spoofing. The modus operandi for phishing works by taking personal information of users from various websites. Sometimes it redirects the user to phish webpage to gain information of user like username, password, account and credit card details etc. The victims are usually asked to enter their information such as account numbers and password. The information entered is taken by phisher and used to gain access to that account. Several tricks are used to fool the user like: duplicating content from official websites, duplicate email addresses looking like legitimate one, advertise on home page that redirects the webpage to phish one.

Phishing is an illicit scheme designed to convince unsuspecting users into giving up their confidential information on falsified sites spoofing as genuine brands. Most of the phishing techniques rely on leading victims to the phishing site by following links in emails, forums, blogs or private chat messages, as well as by clicking on banners representing legitimate sources. Cybercriminals can then harvest credentials submitted at the phishing site to gain access to victims’ account.

In the coming sections, we study how to detect such malicious phishing pages using: Hyperlinks, URL and Terms.

II. PHISHING EMAILS

Phishing email is a common type of attack wherein phishers send out fraudulent emails impersonating genuine service providers and asking the victims to give away their personal information or direct them to bogus websites. The bogus websites look similar to genuine electronic service providers’ website. Once these victims log onto the bogus website, their personal information are recorded by the adversaries. Identity theft is the method used by phishers to acquire credentials of victims for gaining control of their accounts and subsequently embezzling funds out of them. Currently there are several products available in the market that use text classification to limit the potential damage caused by phishing emails. For example, Anti-Phish is a browser extension used to protect inexperienced users against malicious websites. The Anti-Phish plug-in tool keeps track of users’ sensitive information and prevents this information from being passed to a web site that might be considered untrustworthy or unsafe. A text classification algorithm is responsible for identifying whether a given website is a phishing site based on addresses used in a form. It compares a legitimate URL and IP address associated with URL the page locates. Anti-Phish focuses more on tracking sensitive information provided by a user and identifying a website as a suspect phishing site when its visual similarity value is above a pre-defined threshold.

III. PROTECTING PHISHING EMAILS

Phishing emails usually ask the user to click on a hyperlink leading to malicious site. A detailed analysis of phishing emails the phishing hyperlinks can be cast into the following general categories based on their methodologies used [5]:

1) Different actual link and visual link in the email i.e., the hyperlink in the email does not point to the same location as apparently displayed to the users
2) The DNS name in the hyperlink is substituted by the quadruple IP address
3) DNS names used are manipulated to look similar to the genuine DNS names the phishers are trying to forge
4) The hyperlink is encoded to make it difficult to read for example, unusually long hyperlinks
5) When visiting the phishing hyperlink, it usually asks the user for various personal details like username, password, account number, SSN, etc.

Based on the above methodologies, the links can further be classified into:

1) Visible_links: This could be determined by the total number of links in an email.

2) Invisible_links: If the color difference between the background and font of link in an email is less than 500, the link can be considered as an invisible link.

3) Unmatching_urls: A binary value to show whether the visible URL is as the same as the hidden URL.

Values of all the above features are numerical but in a different range. If we find that the value for “Invisible_links” and “Unmatching_urls” is a non-zero value, then we consider the given email as a possible phishing attack. In case a phishing attack is detected the user is notified about the forged email suspicion and is then advised to delete the email.

The algorithm can be summarized as follows [5]:

1) User opens the web browser to open the email on it.

2) The email before opening will be scanned by the backend phishing detection engine.

3) The “visible_links” are to be extracted from the body of the email.

4) The “invisible_links” are then extracted from the email body as well.

5) The “Unmatching_urls” are also extracted from the email body.

6) If the count of “invisible_links” or “Unmatching_urls” is greater than 0, then:

   a. Prompt the user that this could be a phishing attack.

   b. Advise him to delete the mail.

7) Else

   a. Open the email normally.

IV. PROTECTING PHISHING WEBSITES

Phishing website detection system called phisher detector is used which can categorize website as either phishing or legitimate. In following section in detail explanation of system architecture is given which helps to achieve objectives.

Following are the main objectives of the system [8]:

1) Extract terms and URLs from web page using DOM parser.
2) Identify important terms (brand name) using TF-IDF and URL weighting scheme.
3) Search results for brand name using search engine API.
4) Identify victim

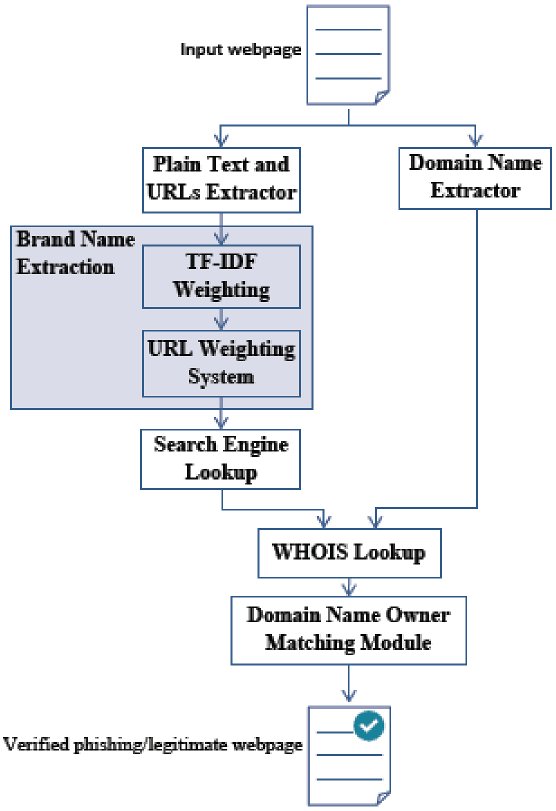

Fig 1: Architecture of Phishing Detector [8]

A. Plaint Text and URLs Extraction: Webpage parsing parses the collected web pages, one page at a time, to get terms and URLs from respective web page . HTML parser is used to create a Document Object Model (DOM). The Document Object Model is a standard for making and controlling in-memory representations of HTML (and XML) content. DOM is presented through tree style structure which is transformable and can be used to reproduce a complete page. The main objective of this part is to extract all portions of plain text that potentially contains the brand name. A brand name is a name given to a legally registered product, service or business and is used as a unique identity.

Here not all the words from the source of web page as plain text. Rather only textual content from specific tags is considered as plain text. List of those tags is as follows:

<meta>...</meta>
<title>...</title>
<body>...</body>.

The URL Extractor can be used to extract URLs from webpages. Generally URLs are placed inside an anchor tag where href property denotes the URL path. Here the anchor tag is parsed. Anchor tag has following syntax structure:

<a href = > ...</a>

<a href = > ...</a>

<a href = > ...</a>

B. Brand Name Extraction: The Shroff word frequency is used to assign weight to words that have been obtained from a web page parser. After generation, the weight is calculated by searching for important keywords in it.

TF-IDF [Term Frequency-Inverse Document Frequency] is a well-known algorithm to find the most important words in a document. The TF-IDF algorithm computes initial weights for all words in the plain text extracted from HTML content. This weight is a statistical measure used to evaluate how important a word is to a document in a collection.

Term Frequency, measures how often does a term occur in a record. Since every record is different in length, it is possible that a term would appear much more time in long records than shorter one. Thus, the term frequency is generally divided by the record length to normalize it:

TF (t) = (Number of times term t appears in a record) / (Total number of terms in the record).

Inverse Document Frequency, measures how important a term is *in the record*. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", "that" etc. may appear a lot of times but are least important. Thus we need to weigh down the frequent terms while scaling up the rare ones, by computing the following:

IDF (t) = log_e (Total number of documents / Number of documents with term t in it)

In URL Weighting System mainly the URL's are weighed according to the occurrences in the document of the source code

C. Search Engine and WHOIS Lookup: At this point, we already have the brand name keywords, but still lacking the knowledge of which legitimate domain they came from. So here brand name keywords are given to a search engine and retrieved the top 30 results. Then analysis of the URL listed at every line of the result is done to obtain its domain name. The domain name with the highest frequency is essentially the domain name belonging to the legitimate website that is being phished. Finally WHOIS lookup is used to compare registration details of websites to correctly categorize website as phishing or legitimate

V. CONCLUSION

In this paper, study of an approach to detect phishing emails using link based features is discussed. Also, to detect a phishing webpage a phishing detector is used which uses url based text extraction feature.

REFERENCES

[1] Patil, Dharmaraj Rajaram, and J. B. Patil. "Survey on Malicious Web Pages Detection Techniques." *International Journal of u-and e-Service, Science and Technology* 8.5 (2015): 195-206.

[2] Tan, Choon Lin, Kang LengChiew, and KokSheik Wong. "PhishWHO: Phishing Webpage Detection via Identity Keywords Extraction and Target Domain Name Finder." *Decision Support Systems* (2016).

[3] Singh, Priyanka, Yogendra PS Maravi, and Sanjeev Sharma. "Phishing websites detection through supervised learning networks." *Computing and Communications Technologies (ICCCT), 2015 International Conference on*. IEEE, 2015.

[4] Eshete, Birhanu. "Effective analysis, characterization, and detection of malicious web pages." *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 2013.

[5] Jain, Aanchal, and Vineet Richariya. "Implementing a web browser with phishing detection techniques." *arXiv preprint arXiv:1110.0360* (2011).

[6] Choi, Hyunsang, Bin B. Zhu, and Heejo Lee. "Detecting Malicious Web Links and Identifying Their Attack Types." *WebApps*. 2011.

[7] Senavirathne, W. M. N. A Machine Learning Based Approach for Phishing Detection On Smart Phones. Diss. 2016.

[8] Tan, Choon Lin, Kang Leng Chiew, and KokSheik Wong. "PhishWHO: Phishing webpage detection via identity keywords extraction and target domain name finder." *Decision Support Systems* 88 (2016): 18-27.

[9] Tan, Choon Lin, and Kang Leng Chiew. "Phishing website detection using URL-assisted brand name weighting system." *Intelligent Signal Processing and Communication Systems (ISPACS), 2014 International Symposium on*. IEEE, 2014.

[10] https://webcache.googleusercontent.com/search?q=cache:bx18zXz5-vIJ:https://digitalguardian.com/blog/social-engineering-attacks-common-techniques-how-prevent-attack+&cd=1&hl=en&ct=clnk&gl=in