Website Security for Detection of Phishing Sites

Supriya S. Shinde, Kirti K. Singh, Chaitali A. Patil
Department of Computer Engineering,
K.C. College of Engineering & Management studies
Kopri , Thane(E)-400 603, India.

Abstract— The word "phishing" comes from the analogy that Internet scammers are using fake email to steal for Passwords and personal financial data from the sea of Internet users. In this paper, we present the design, implementation, and evaluation of CANTINA, a novel, content-based approach to detecting phishing web sites. We also discuss the design and evaluation of several heuristics we developed to reduce false positives. Our experiments show that CANTINA is good at detecting phishing sites, correctly labeling approximately 95% of phishing sites. In this project we are going to use features for detect phishing sites through automated whitelist, inspection of logos and page layout and also hyperlinks detection.

Keywords—Phishing,OnlineBankingtheft,Automated Whitelist,Blacklist,Url Detection.

I. INTRODUCTION

Phishing is a form of social engineering attack used by cyber criminals to steal sensitive information. Customers of leading Banks throughout the world have been a target of Phishing. Our project focuses on the security measures that financial service providers such as Banks can take to prevent and manage a Phishing attack. Most Phishing attacks use a combination of fake emails and look-alike websites to fool the users into revealing their personal financial details. Users are usually sent an official looking forged email that appears to come from the genuine organization but is actually sent by the attackers. This email lures the users into visiting a fake website where they logon and update their personal information there by revealing their details to the attackers.

Fig 1.1 Basic Idea Of The Project

A.Purpose of this project

Banks manage sensitive information of their customers, both personal as well as financial details. Apart from technical controls in the web application, it is important for such organizations to follow secure processes while handling any customer information. Secure internal processes would help in preventing any leakage of customer information including email addresses that may be used for Phishing. This can include activities such as*:*

1. Restrict customer database access to authorized users only.
2. Dispose media only after erasing the data containing user information if any.
3. Make all the personnel handling customer data aware of confidentiality requirements and the risks of breach.
4. Do not display email id in any mass mailers.
5. Share email addresses only with authorized marketing alliances or other groups with similar security controls.

B. Existing System

The exponential growth in online financial transactions has made Phishing a lucrative option for attackers. Today almost all the banks provide online banking facilities and the customers of these banks can easily become a target of Phishing. Using stolen information attackers can perform a number of fraudulent activities, which may include:

1. Carrying out unauthorized transactions using credit or debit card numbers.
2. Logging into the banking application using username and passwords. The attacker can get access to all the financial details of the user, as well as conduct transactions on his behalf.
3. Selling user's personal information such as phone numbers, address, account numbers etc to others for different mischievous activities.
4. Denying service to legitimate users by changing passwords and other contact details.
5. Ruin the customer's trust in the services provided by the bank and malign the brand name.

II.LITERATURE SURVEY

There are certain technology that already exists in the area of detecting phishing sites. Some of them are as mentioned. Each of them deals in the similar way the only differentiation is its technique in recognizing phishing sites. The following are the different approach that were implemented in solving the phishing problem.

Textual and Visual Content-Based Anti-Phishing: A Bayesian Approach IEEE paper was published in October 2011 presented a novel framework using a Bayesian approach for content-based phishing web page detection is presented. This model takes into account textual and visual contents to measure the similarity between the protected web page and suspicious web pages. A text classifier, an image classifier, and an algorithm fusing the results from classifiers are introduced. An outstanding feature of this paper is the exploration of a Bayesian model to estimate the matching threshold.

Cantina+ : Feature-rich machine learning framework for detecting phishing web sites IEEE paper published in 2010 which presents specifically, we proposed CANTINA+, the most comprehensive feature-based approach in the literature including eight novel features, which exploits the HTML Document Object Model(DOM), search engines and third party services with machine learning techniques to detect phish. Moreover, we designed two filters to help reduce FP and achieve runtime speedup. The first is a near-duplicate phish detector that uses hashing to catch highly similar phish. The second is a login form filter, which directly classifies webpages with no identified login form as legitimate.

III. PROPOSED SYSTEM

According to the literature survey, we have come across many limitations which we are trying to overcome in our proposed system. Our system is proposed to overcome the drawbacks mentioned in the existing system and try to meet as much points mentioned below:

As it is seen that previously phishing detection is only available for specific websites (eg.Ebay & Google), this system will focus on **"**All Banking Websites".In this system we focus on the content based approach i.e. CANTINA for retrieving of information of phishing websites .TF-IDF is an algorithm often used in information retrieval and text mining. The drawbacks of system can be overcome through advanced features such as Automated Whitelist, Inspections Of Logos, WHOIS Information, Inspection Of Html Content, Automated Blacklist,IP Address Properties.

A. Specifications

- **Hardware:**

a. Processor: Pentium 4

b. RAM: 512 MB or more

c. Hard disk: 16 GB or more

d. Android Device

- **Software:**

a. WAMP server

- **Frontend:**
  a. PHP
- **Backend:**
  a. MYSQL

1.Methodology

- In this system will crawl the original site of bank and it will retrieve all urls, location of bank's server and WHOIS information.
- If user get any email with phishing attack link.
- Then our system will take that url as input and crawl the link, retrieve all urls .
- And system will compare these urls with original banks url database, try to find urls are similar or not.
- Then system will find location of Phishing link url and compare location with original banks location.
- After that system will find out Whois information of URL.
- System will analyze the information and show the results to the user.

The flowchart diagram of the system is given in figure 4.1.

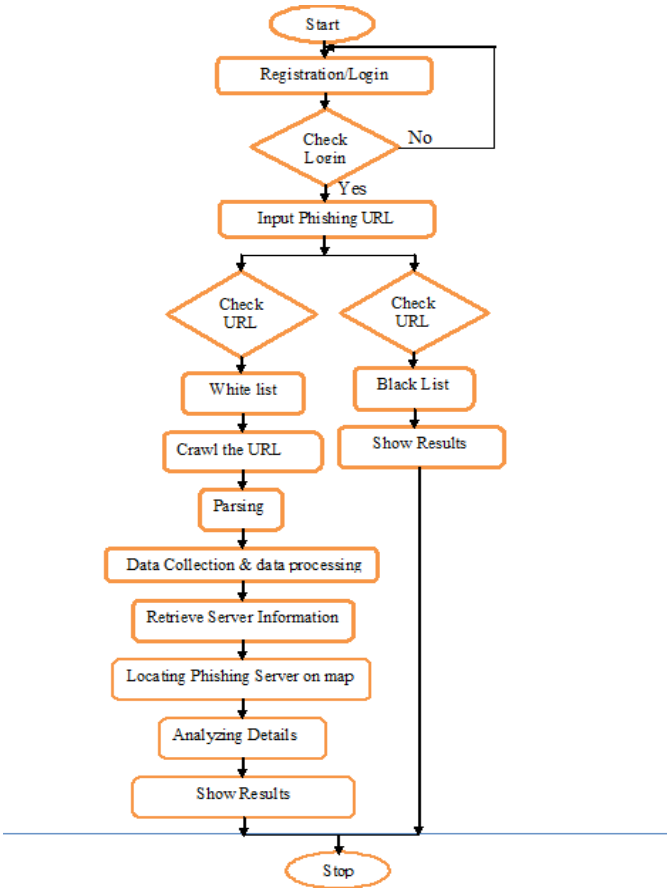

Fig 3.1 Flowchart Of The Project

2.*ALGORITHM*

TF-IDF is an algorithm often used in information retrieval and text mining. TF-IDF yields a weight that measures how important a word is to a document in a corpus. The importance increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus. The *term frequency* (TF) is simply the number of times a given term appears in a specific document. This count is usually normalized to prevent a bias towards longer documents (which may have a higher term frequency regardless of the actual importance of that term in the document) to give a measure of the importance of the term within the particular document. The *inverse document frequency* (IDF) is a measure of the general importance of the term. Roughly speaking, the IDF measures how common a term is across an entire collection of documents. Thus, a term has a high TF-IDF weight by having a high term frequency in a given document (i.e. a word is common in a document) and a low document frequency in the whole collection of documents (i.e. is relatively uncommon in other documents).

B. Working

- In our system we are checking the URL and domain name of the particular bank website.
- Age of Domain Name: The domain names used by fraudsters are usually used for a limited time.
- Presence of Form Tag: HTML forms are one of the techniques used to gather information from users.
    - Automated Whitelist
    - Automated blacklist.
    - Inspections of logos.
    - URL is nothing but IP Address.
    - Using IP address our system will locate phishing server.

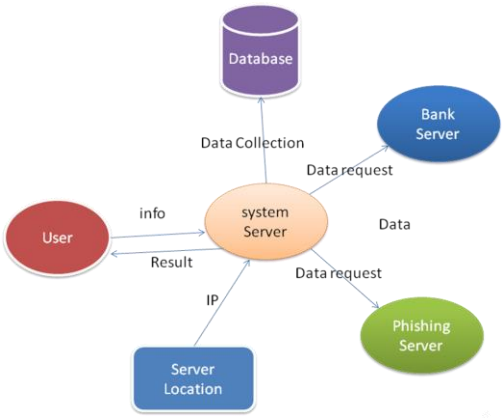

Fig.3.2.Architecture of the system

IV. CONCLUSION

The problem of Phishing does not have a single solution as of now. Phishing is not just a technical problem and Phishers would keep coming up with new ways of attacking the users. Banks should undertake periodic vulnerability analysis to identify and plug weaknesses that can lead to a successful Phishing attack. The solution lies in a combination of controls setup by the organization and user awareness.

REFERENCES

[1] Sheng, S.,Wardman, B.,Warner, G., Cranor, L., Hong, J., and Zhang, C. 2009. An empirical analysis of phishing blacklists.
[2] eBay Toolbar's Account Guard (2011). <http://pages.ebay.com/help/confidence/account-guard.html>.
[3] Xiang, G. and Hong, J. 2009. A hybrid phish detection approach by identity discovery and keywords retrieval. In *Proceedings of the 18th International Conference on World Wide Web (WWW'09)*. 571–580.
[4] Xiang, G., Pendleton, B. A., Hong, J. I., and Rose, C. P. 2010. A hierarchical adaptive probabilistic approach for zero hour phish detection. In *Proceedings of the 15th European Symposium on Research in Computer Security (ESORICS'10)*. 268–285.
[5] Zhang, Y., Hong, J., and Cranor, L. 2007. Cantina: a content-based approach to detecting phishing web sites. In *Proceedings of the 16th International Conference on World Wide Web (WWW'07)*. 639–648.
[6] Aburrous, M., Hossain, M. A., Dahal, K., & Thabtah, F. (2010). Intelligent phishing detection system for e-banking using fuzzy data mining. Expert Systems with Applications, 37(12), 7913–7921.