Intelligent Detection of Phishing E-banking Website Using Fuzzy Datamining

Nataasha Raul, Chinmayee Vaidya, Pooja Kolhe, Khushbu Nehita

***Abstract*—Recently there has been an increase in the number of phished websites. Numerous approaches are adopted by phishers to conduct a well-planned phished attack. The victims of the phishing attacks, are mainly the on-line banking consumers and the payment service providers. They have to face substantial financial loss which eventually results in lack of trust in Internet-based payment and banking services. In order to overcome these huge loses there is an urgent need to find solutions to in order to combat the attack of phishers. The detection of phished website is extremely complicated and requires outstanding expert experience and knowledge. Till now, numerous solutions have been proposed in order to address these problems. Most of the solutions developed have failed to make a dynamic decision on whether the site is phished or not, which has eventually lead to a number of false positives. In our research we developed an application of an intelligent fuzzy logic based system for e- banking phished website detection. The most important aim of our proposed system is to protect the users from deceitfulness of the phishers, and helping them to detect whether the website is safe or phished**

Index Terms—Phished, fuzzy, features, fuzzified

I. INTRODUCTION

Phishing is an act where fake websites attempt to obtain users credentials. These days sophisticated phishing acts are prevalent all over the world. In E-Banking phished websites different websites are faked by malicious hackers who mimic real websites to obtain User information. Some of these websites closely remember the real ones. Unknowingly users may submit their valuable details in phished websites and may be easily fall prey to this scam. Victims may expose their important details such as user name password, credit card number and account number to the owners of phished websites. This impact is a breach of information security due to the compromise of confidential data and the victims may have to bear serious financial losses. Compared to other cyber-crimes i.e. virus and hacking Phishing is a new kind of internet crime. Our approach towards this problem is the use neuro-fuzzy systems using Legitimate Rules, user behaviour Profile, Phish Tank, User Specified Site and pop- up windows from e-mails.

Manuscript received April 28, 2014.

Prof.Nataasha Raul, Assistant Professor,Computer Department Sardar Patel Institute of Technology, Mumbai, India, (**e-mail- nataasharaul@spit.ac.in)**
Chinmayee Vaidya, Computer Department, Sardar Patel Institute of Technology, Mumbai, India (e-mail- vchinu25@gmail.com),
Pooja Kolhe, Computer Department, Sardar Patel Institute of Technology, Mumbai, India, (email-poojakolhe2193@gmail.com),
Khushbu Nehita, Computer Department,Sardar Patel Institute of Technology, Mumbai, India, (e-mail-khushbunehita@gmail.com)

II. RELATED CONCEPTS

A. Extracting features

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 figure1 which would be used by us in our proposed system.[1] The list given below demonstrates all the 20 collected phishing website features, which will be used by us later in our proposed methodology for our fuzzy-based phishing detection model: [1]

1. Spelling errors: Most phished websites have spelling and grammatical since they are created on a temporary basis and the phishers are always in a hurry. Increased number of spelling errors could be a sign of phishing website.

2. Long URL address: A website having a short URL address is much more trustworthy and reliable than a website with very long URL address. For eg. Website with a short URL address like"http://www.cman.com"is much more reliable than the website with URL address:"http://www.cman.ViewOver?DocId=Index1 siteId=A C

3. Using SSL certificate and padlock icons: The website which possess secured encryption transaction SSL certificate i.e. https:// can be considered more trustworthy and reliable than the unsecured website i.e. http://. Most of the phished websites don't use this feature for numerous reasons. Hence the usage of SSL can be distinguished by taking into consideration the padlock at the bottom of a browser frame.
4. Certification authority: The existence of mouse-over can be considered as a fake certification authority and digital signature. This feature can be detected using the source code of the website.

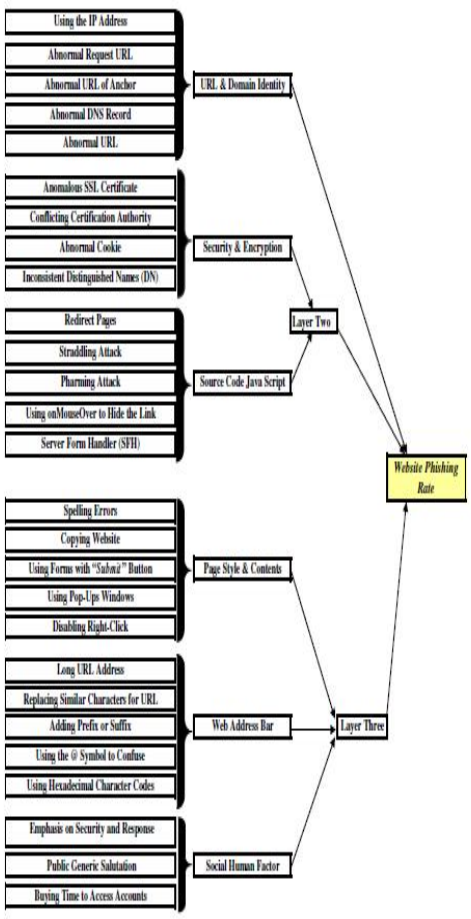

Figure 1: Phishing Detection Features [1]

5. Replacing similar characters for URL and registered domains: Converting the real URLs by replacing characters such as an uppercase I with a lowercase L or the number with 1 or uppercase S by 5. For eg. Transforming www.citi.com to www.cit1.comand also using and registering a domain name which is very similar to the original one, and is owned by a respectable company.

6. Adding a prefix or suffix: Adding a prefix or suffix to the real domain name, such as www.citibank.com with www.online-citibank.com or with www.online-card.com. For example, we can see the usage of prefix word "online" before the legitimate Citibank domain name, in order to obscure the user.

7. Redirect pages: Utilizing programming bugs in real websites in order to redirect to other pages. For example, the Citibank site used to include a script that could redirect users to any site specified in place of PHISHED link In the URL: http://citibank.com/ws/citibankISAPI.dll?MfcISAPIComman=RedirectToDomainDomainUrl=PHISHING LINK.

8. Straddling attack. The owners of phished websites insert spiteful data in the HTML code of the long-range web page. When the web page is downloaded, the hidden script inside the HTML page will be executed.

9. Using forms with Submit button: The usage of submit button at the bottom of the form causes the information to be deceptively sent to the fraudster's specified location. For eg. < FORM action =http://www.axisoffer.com/sendmail.php method=get target= blank >

10. Using onMouseOver to hide the Link: Using JavaScript event handler onMouseOver to show a false URL in place of status bar.

11. Using the IP Address: Fraudsters attempt to hide the destination website by concealing the URL. One method of hiding the destination address is to use the IP address of the website, instead of its actual hostname. Here is an example of an IP address used in a fraudulent website: http://210.15.220.65/sr/.

12. Using the @ Symbol to Confuse: When the@ symbol is used in the URL, all text that comes before the @ symbol is ignored and the browser takes into consideration only the information following the @ symbol. In other words, if the format <user info>@<host name>is used, the browser is directed to the <host name>site and the part<user info>is ignored. In order to conceal the URL even further, the@ symbol can be replaced by its hexadecimal character code i.e. %40.

13. Using Hexadecimal Character Codes: Phishers can also conceal URLs by using hexadecimal character codes in order to represent the numbers in

the IP address. Each of the hexadecimal character code begins with % sign. Eg. http://www.visa.com%00@%33%33%17%3E%40%37%3E%22 %21%35%3E%33%33%32 Here the URL is represented in <user info><null>@<host name>format.

14. Disabling Right-Click: Usage of Java Script in order to disable the right-click function, which prevents the user from viewing and saving the source code. Sometimes the right-click function is also disabled in phished webpages that are opened in the menu browser window. The following is JavaScript taken from a fraudulent PayPal website. Function click () if (event.Button==2) alert (WARNING! Copyright 1999-2004 PayPal. All Rights Reserved.).

15. Buying Time to Access Accounts: Phishers try to buy some time before the victims submit their valuable credentials to their accounts to give the phishers a chance to make use the information they have acquired from the victims. The scammers indicate that the web pages will require a certain amount of time for the updating the account. They just hope that this will avert their victims from checking their accounts during this period.

16. Abnormal Request URL (RURL): External objects like css, images, and external scripts in a web page get loaded from another URL. For websites such as, a large portion of those URLs are in its own domain.

17. Abnormal URL of Anchor (AURL): A web page can be considered suspicious if the domains of most of the AURLs are differ from the pages domain, or anchors that do not link to any other page. A high portion of those anchors in a legitimate website usually point to the same domain as the page. (For eg: <ahref=http://www.axisbank.com/>).

18. Abnormal URL: If the host name in URL does not match with its claimed identity it can be considered as an instance of abnormal URL.

19. Server Form Handler (SFH): Most of the e-banking websites contains a server form handler. For phished sites, the SFHs are usually like about: blank or they refer to a completely different domain.

20. Abnormal Cookie: In a phished website, the cookies point to its own domain, which can be considered inconsistent with the claimed identity of the website. They may also point to the real site, which is also inconsistent with its own domain.

B. Fuzzification

This is the process of generating membership values for a fuzzy variable using membership functions. The first step is to take the crisp inputs from the 20 characteristics and factors which resemble the fake phished website and determine the degree to

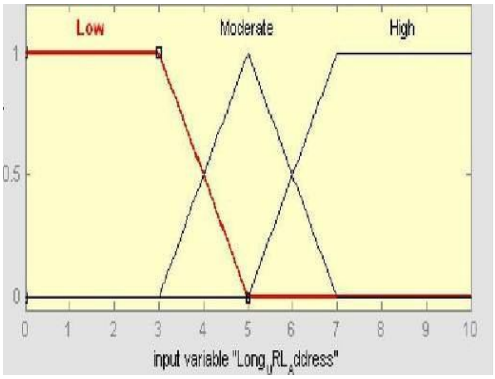

Figure 2: Membership Function of a long URL address

URL Address Length – Low, Moderate, High.
Linguistic Variable: URL Address Length [1]
Linguistic value Numerical Range
Low [0, 3, 3, 5]
Moderate [3, 5, 7]
High [5, 7, 10, 10]

which these inputs belong to each appropriate fuzzy set [1].This crisp input is always in the form of a numeric value which is limited to the universe of discourse. Once the crisp inputs are obtained, they are fuzzified against the appropriate linguistic fuzzy sets. The fuzzy detection model provides us definitions for each factor and its interactions with various other factors. This approach will provide a decision tool for identifying phishing websites. The advantage offered by fuzzy logic techniques is the use of linguistic variables to represent key phished website characteristic indicators. Here descriptors such as low, medium and high are assigned to a particular range of values for each of the key predefined phished website characteristic indicator. Since these descriptors form the basis for capturing expert inputs based on the impact of predefined phishing characteristic indicators on the Phished Website, and it is important to calibrate them so that

they can be interpreted by the experts providing input. The valid ranges of each of the inputs are considered and are divided into fuzzy sets. For eg. URL length can range from low to high along with other values in between. The degree to which each of the values belong to each of the fuzzy sets is called membership function. A membership function is then designed for each phished website characteristic indicator, which is a curve that determines how each and every point in the input space can be mapped to a particular membership value or the degree of membership between 0 and 1. The linguistic values can be then assigned to each of the phishing indicator as low, medium and high and for the extent to which a website is phished asLegitimate,Suspicious and Phished (triangular and trapezoidal membership function).For eg. Linguistic descriptors can be used to represent one of the important phished website features such as URL address length and a plot of the fuzzy membership functions is shown in Figure 2 below. Each plot has x-axis which represents the range of possible values for the corresponding key phishing characteristic indicators (Low, Moderate and High) and the y-axis represents the degree to which a value for the vital phishing characteristic indicators is represented by the linguistic descriptor. For example, the plot of the membership function for URL Address Length, 4.5 cm is considered Low with a membership of 30% and is also considered Moderate with a membership of 65%. The fact that 4.5 cm URL Address Length is considered both Low and moderate to varying degrees is a distinguishing feature of fuzzy logic. The same approach can be used to calibrate all the other key phished website characteristic indicators

C. Defuzzification

This is the process in which a fuzzified output of a fuzzy inference system is converted into a crisp output. Fuzzification aids us to evaluate the rules, but eventually the output should be in the form of a crisp number. Now, the input given for the defuzzification process is the summation of the output of the fuzzy set and the final output is in the form of a number. The output helps us to determine the risk rate of the website under consideration i.e. Legitimate, Phished, Very- Phished.[1]

III.IMPLEMENTATION

A.Proposed System Architecture

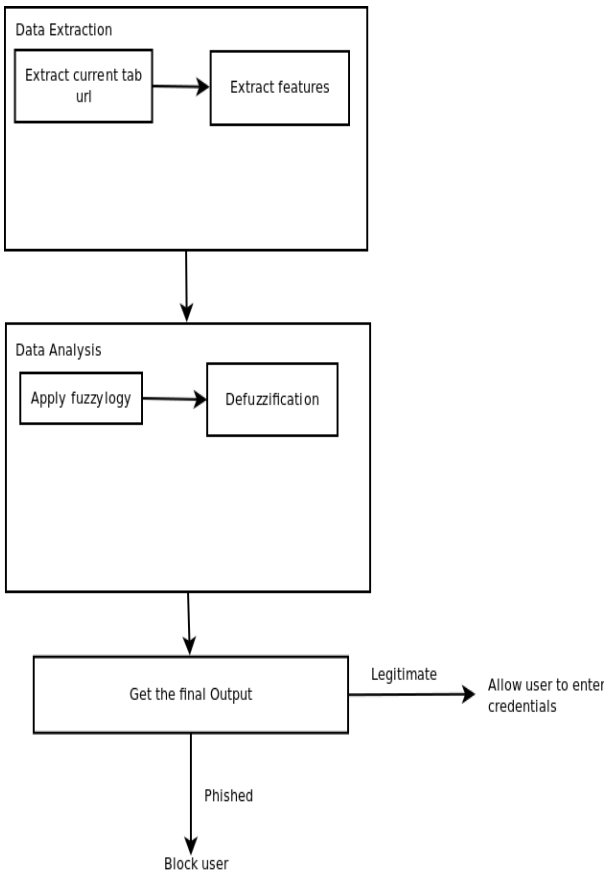

Figure 3: Our Proposed system architecture

B.Our Proposed System

We have developed an application for the detection of phished e-banking websites using C# platform. We have created a simple web browser from where we will extract the URL before the page gets loaded. On the extracting the URL, we will be detecting the presence of 20 features and then applying fuzzy logic to get the final output .These all operations we are applying before the page gets loaded. [6] [7] If a website is safe then it will get loaded, but if a website is suspicious then warning will be displayed to the user saying that the website may be phished and if the website is phished we will not allowing the website to load .We are going to use the predefined phished website features mentions above. We will get a value between 0-5 for each of the features based on the extent to which it is phished. The values allotted can be divided into three parts:

- Range 0-1.5- Legitimate
- Range 1.5-3.5-Suspicious
- Range 3.5-5-Phished

The values now obtained will be given as input to the Mamdani Fuzzy Controller. The fuzzy controller now has 20 inputs obtained because of 20 features. Here, we have designed a simple fuzzy controller, where the 20 inputs are fuzzified using IF-THEN rules and then later defuzzified using centroid method to obtain a crisp value. The crisp value helps us to determine whether the website is legitimate, suspicious or phished. If a website is detected as phished it will be added to the database, so that next time if the same website is entered, it will be first checked in the database and if it is already declared as phished, the site will be blocked.[2] For the feature URL length, we are going to use the following logic: If URL length<57 Safe Else IF URL length>=57 and <75-Suspicious Else IF URL length>=75-Phished [2] For detecting websites where hexadecimal characters are used we are just going to convert the hexadecimal codes in the URL into their corresponding characters and then by taking into consideration what these characters represent we will be able to determine whether the website is phished or not. For features such as using mouse-over to hide link, disabling right click and redirect pages we have used the source code of the website.

For eg. If the anchor tag in the website points to some other website than it was supposed to we can say that it is a phished website. For features such as using @ to confuse and using prefix and suffix we are just going to detect the presence characters such as @ and - using toString() function. [2] Eg. If URL has @ symbol-Phished Else -Not phished. If URL has - symbol- Phished Else Not phished

For the feature replacing similar characters for URL and registered domains we have created a database of all banks and their websites. In that database corresponding to every bank name we have stored its website and also its domain name. So the domain name of the URL is extracted and it is checked with the domain name in the database according to the URL entered. If the domain name matches with the one stored in the database, then we can say that there is no spelling error or else we can say that some characters in the website's URL is replaced with similar characters. For all the domain based features we have used whois.com.

DOMAIN	SERVER
vc	whois2.afilias-grs.net
ve	whois.nic.ve
vg	whois.adamsnames.tc
ws	whois.website.ws
xxx	whois.nic.xxx
yu	whois.ripe.net
za.com	whois.centralnic.com
ac	whois.nic.ac
ae	whois.aeda.net.ae
aero	whois.aero
af	whois.nic.af
ag	whois.nic.ag
al	whois.ripe.net
am	whois.amnic.net
as	whois.nic.as
asia	whois.nic.asia
at	whois.nic.at
au	whois.aunic.net
ax	whois.ax
az	whois.ripe.net
ba	whois.ripe.net

Figure 4: whois.com database

We have used its database which consists of domain names and its corresponding server names. So according to the domain in the extracted URL the corresponding server name is chosen. This server name is then entered into the whois.com API which later extracts all the domain data from the whois database. This data then helps us to perform operations on the domain related features. [3]

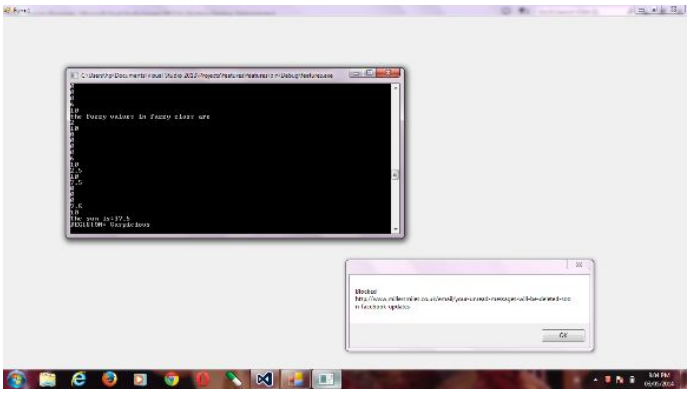

Figure 5: Phished website detected and blocked by our system.

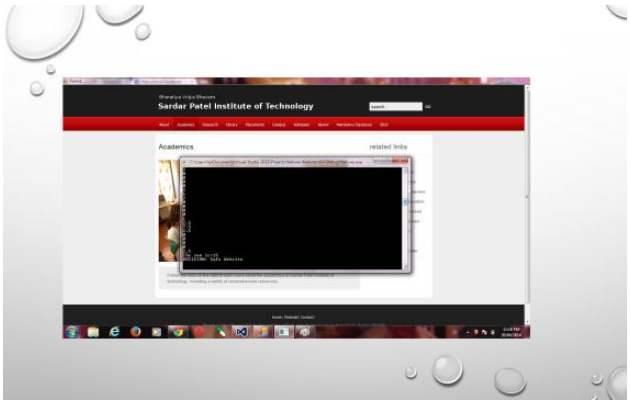

Figure 6: Safe website detected by our system and allowed to get loaded.

IV. FUTURE SCOPE

In the future we are planning to use MATLAB for graphical representation. We are also planning to use alexa.com to introduce web traffic analysis for the detection of phished websites. Introduction of web traffic analysis will help us increase the accuracy of the detection of such websites. We will also try to reduce the number of false positives and false negatives.

V. CONCLUSION

Fuzzy logic combined with the analysis of predefined phished website features can prove to be an efficient way for building intelligent model for the detection of phished websites. Hence detection of phished websites will be more accurate and people will be averted against submitting their valuable credentials to those phished websites.

REFERENCES

[1] Design and Development of an Intelligent Association Classification Mining Fuzzy Based Scheme for Phishing Website Detection with an Emphasis on E-Banking by Maher Ragheb Mohammed Abur-rous Sub- mitted for the degree of Doctor of Philosophy Department of Computing University of Bradford (2010)

[2] The 7th ICITST 2012 Conference -An IMPORTANT Assessment of Features Related to Phishing Websites using an Automated Technique (2012)

[3] Anomaly Based Web Phishing Page Detection by Ying Pan, Xuhua Ding, School of Information Systems, Singapore Management University (2006)

[4] WHOIS,[Online]. Available: http:\\who.com\ [Accessed: 01 Mar. 2014]

[5] PhishTank, [Online]. Available: http:\\www.phishtank.com\. [Accessed: 25 Jan. 2014]

[6] Fuzzinator: A fuzzy logic controller, [Online]. Available: http:\\www. codeproject.com\Articles\33214\Fuzzinator-A-Fuzzy-Logic-Controller. [Accessed: 08 March. 2014]

[7] Fuzzilite, [Online]. Available: http:\\www.fuzzylite.com\. [Accessed: 08 March. 2014]

[8] Websites of Banks in India, [Online]. Available: http:\\www.rbi.org.in\ scripts\banklinks.aspx. [Accessed: 08 March. 2014]

Prof .Nataasha Raul has B.Tech and Masters Degree in the field of Information Technology . She works as an Assistant Professor at Sardar Patel Institute of Technology, Andheri (W). Her areas of interest include computer programming, data structures, and analysis of algorithm, web engineering and operating systems. She has a huge body of research work under her.

Ms.Chinmayee Vaidya is currently studying Bachelor of Engineering in Computer Science at Sardar Patel Institute of Technology. She worked as the Head of Events of Computer Society of India (C.S.I) branch of the college and organised various technical and non-technical workshops for the students of the department. She has a good academic record and her areas of interest include Computer Networks, Information Security, artificial intelligence and soft computing.

Ms.Pooja Kolhe is currently studying Bachelor of Engineering from Sardar Patel Institute of Technology. She is mainly interested in research work and her areas of interest include Datamining and operating systems.

Ms.Khushbu Nehita is currently studying Bachelor of Engineering from Sardar Patel Institute of Technology and her areas of interest include databases and information security.