Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 4, April 2014, pg.853 – 861

RESEARCH ARTICLE

BIRCH and DB-scan Techniques in Phishing and Malware Detection

Shabin Blesson[1], Dr. R.Ravi[2], Dr. Beulah Shekar[3]

[1]PG Scholar, Department of Computer science and Technology, Francis Xavier Engineering College, India
[2]Professor, Department of Computer science and Technology, Francis Xavier Engineering College, India
[3]Associate Professor, Department of Criminology and Criminal Justice, Manonmaniam Sundaranar University, Tamil Nadu, India

shabin1988@gmail.com [1], csehod@francisxavier.ac.in [2], fxhodcse@gmail.com [3]

Abstract— Malware and phishing detection is one of the most fascinating topics in recent era because of the harm produced by them to the internet users. Phishing website detection can be said as new to the arena. Phishing websites are considered as one of the lethal weapon to embezzle one's personal information and use it for the crackers benefits. In spite of the fact that malware samples and phishing websites share common attributes, they are created and unleashed to the common world in thousand per day. Since its entry to the internet world, detecting the phishing websites and malwares samples is been a tremendous test to the internet security experts. Many clustering techniques have been deployed to tear apart the phishing websites and the malware samples. The Detection course has been divided into two steps 1) Feature Extraction, where the ideal features are extracted to capture the nature of the files samples and the phishing websites. 2) Categorization, where exceptional techniques are used to automatically group the file samples and the websites into different classes. In this paper, we develop an automatic categorization system to class Malware samples and phishing websites using a cluster ensemble by combining the clustering solutions provided by different base clustering algorithms.

Keywords— Malware samples; phishing websites; Feature Extraction; Categorization; Base Clustering algorithms

1. INTRODUCTION

Phishing is a kind of integrity theft that occurs when a malicious webpage masquerade as the legitimate one to gain sensible information about the users password, credit card number or the account details. Though there are many anti phishing software's and techniques for detecting probable phishing attempts, crackers come up with new techniques to bypass the available software and techniques. Phishing is a trickery technique which employs social engineering and technology to gather sensitive information, such as credit card numbers, passwords, and account details by masking as a trustworthy person or business in an electronic communication. Phishing makes use of delude emails that are made to look authentic and pose as coming from a legitimate user such as financial institution, ecommerce sites etc., to decoy users to visit the fraudulent websites through the link provided in the email or the webpage. The fraudulent webpages are designed to look alike the original page of a real company. The crackers hoax users by employing different social engineering tactics such as to suspend user accounts if they do not complete the account update process provide other information to validate their accounts or some other reasons to get the users to visit their spoofed web pages. Is it important to tackle the problem of phishing? According to the Anti-Phishing Working Group, there were 45,115 unique phishing sites reported in September 2013. Phishing sites mostly targeted the payment services and financial services. It is reported that 56.30 % emails come from phishing websites are related to payment services. Also Phishing websites have targeted 379 brands to show themselves as legitimate as of Anti-Phishing Working Group's report on September 2013.

Malware stands for malicious software which is used to disturb computer operation, collecting sensitive information or gaining control over a private computer. It can be in the form of scripts, live contents and other software. Malware is a term used to indicate invasive software. In all countries it is a serious offence to create malware and distribute malware, but the crackers continue to produce malware for making money. Malware includes computer viruses, ransom ware, worms, Trojan horses, rootkits, key loggers, spyware, adware, malicious BHOs, rogue security software etc., the most of the live malwares are worms and Trojans comparing with viruses. The law says, malware is known as a computer pollutant, as mentioned in the legal codes of U.S. government. Malware is different from deficient software, which is genuine software but contains damaging bugs that were not fixed before release. However, some malware is known as genuine software, and may come from an authoritative company website in the form of an attractive program which has the adverse malware included in it.

Clustering can be considered as the most unsupervised learning problem, it deals with locating a structure in a collection of unlabelled data. A definition of clustering is "an action of constructing objects into groups where its properties are similar in any way". A cluster is therefore said as a collection of similar objects. The ideal goal of clustering is to resolve the peculiar grouping in a set of unlabelled data. A good clustering can be found if there is no absolute criterion.

2. SYSTEM DESIGN

Malware samples and Phishing websites shows common features and the malware samples and phishing websites have an interconnection between them. Through a Phishing link we can send a Malware which can be downloaded when a user clicks it and infect his/her system and steal his identity. Also Phishing website alone can steal information from users by redirecting the webpage to multiple webpages and then returning the page to the default webpage. In the system design, different clustering techniques has been designed and implemented. Here we use BIRCH (Balanced iterative reducing and clustering using hierarchies) and DB-Scan algorithm.

2.1. Proposed System

The basic clustering techniques are changed in the proposed system. The clustering techniques used are BIRCH and DB scan where BIRCH stands for balanced iterative reducing and clustering using hierarchies. The change in the basic clustering technologies provides that this system can be used in larger space. As the phishing websites and malware samples are increasing rapidly the ACS has to capture them and store in the database for the further use. So it needs a very large data sets which can be provided by the BIRCH and DB scan. The results are also very good compared to the existing system as it detects the maximum number of the phishing websites and malware samples. The malware samples and phishing websites are taken from the internet.

2.2. System Architecture

The entire architecture of the working system is shown here. This Simple architecture shows that the user gives the input in the address bar such as the webpages or click on some link shown in some particular webpages. As the user gives the input, it is processed in the number of clusters stored in the database and then shows the user whether the webpage given or clicked is a legitimate webpage or not. If not, the user gets the message as the webpage is a phishing website or the content contained in the webpage is a malware.

This entire mechanism works under the principle of the clustering techniques and the clustering techniques used are the BIRCH and DB-scan which can be used in large datasets as it is a greater advantage in this internet world.

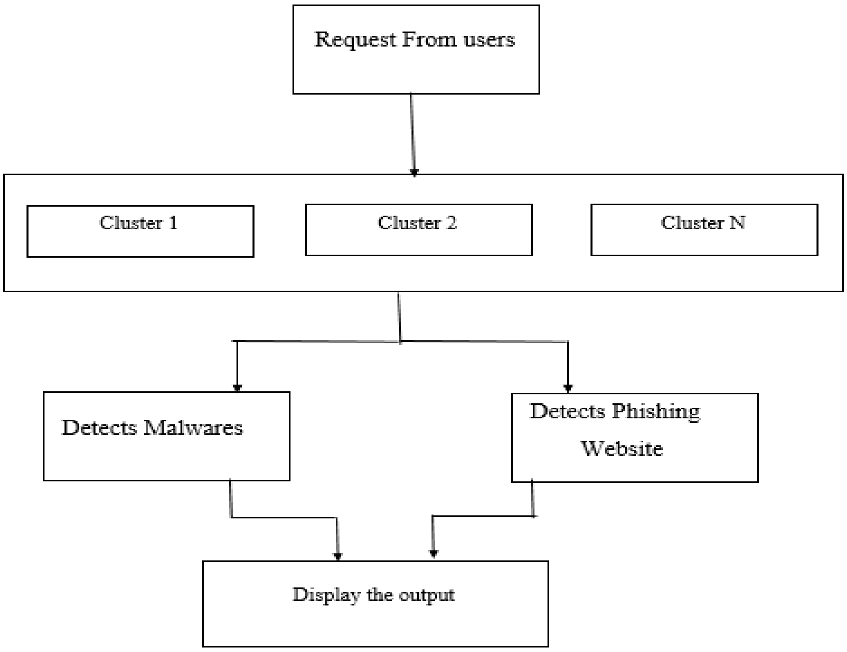


Fig 1 System Architecture

3. SYSTEMIMPLEMENTATION

3.1 Feature Extractor Module

In the feature extraction module we are segregating the malware samples and phishing websites by means of features present in the samples or the websites and we store the data in the database after converting it into a 32-bit file. Then this step moves on to the next step. The connection between databases is established.

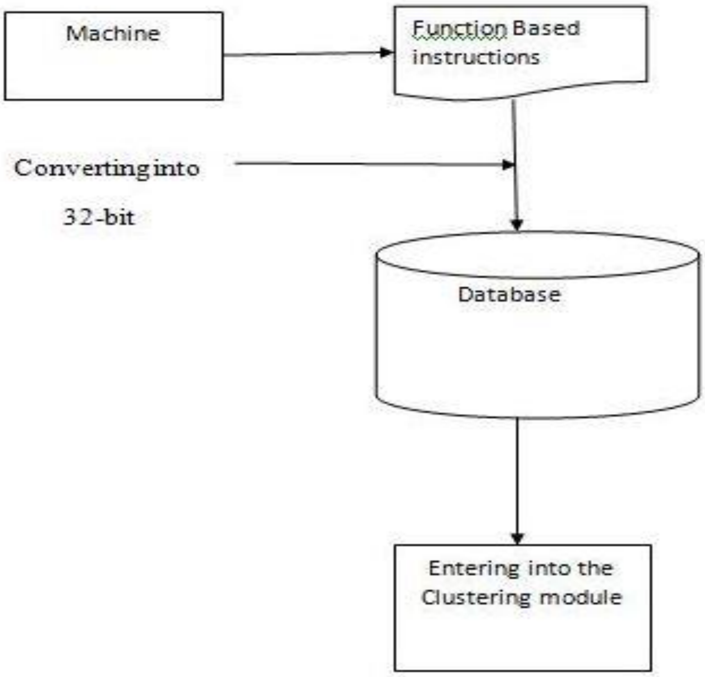

Machine
Function Based instructions
Converting into 32-bit
Database
Entering into the Clustering module

Figure 2 Feature Extraction Module

3.2. Clustering analysis Module

In the clustering analysis module, the 32 bit files fetched from the database are then separated as if they belong to BIRCH clustering or DB scan. If the stored files is in the form of terms then it goes to BIRCH clustering and if it is in instructions then it moves to DB scan.

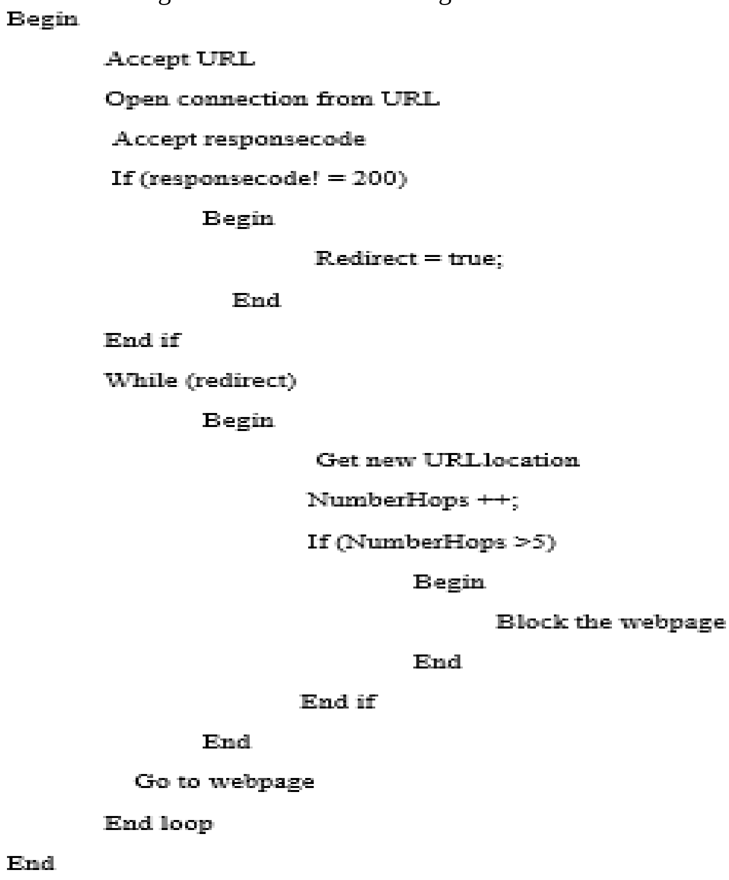
3.3. Black-List Analysis

In Black list analysis, the black listed web pages and malware samples features will be stored in the database which provides the information about the particular samples and the websites such that, whether the website is a phishing suspected and whether the sample contain malicious content that may harm the users computer.

3.4. Blocking access

In this module the Message is displayed as the malware sample or the phishing website is a threat and so it allows the user to block the particular webpage or malware.

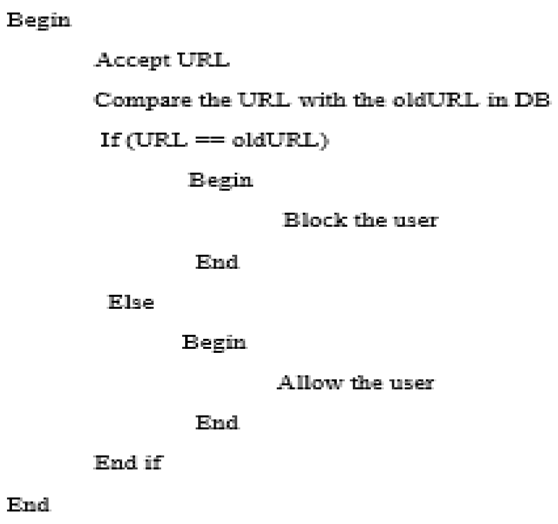
This is the System Modules and the algorithm of this process is given below. In this algorithm, the links given are analysed. The URL connection is established if the redirects code is satisfied. If not it is considered as phishing website and displayed to the user as such. The redirect works such as the number of redirects happens to that URL when the URL is given as the source. If the URL crosses the certain limit then it is considered as a suspicious URL. This algorithm is followed when a link is clicked from the email are somewhere in the internet. The algorithm for the redirect is given below.



```
Begin
        Accept URL
        Open connection from URL
        Accept responsecode
        If (responsecode! = 200)
                Begin
                        Redirect = true;
                  End
        End if
        While (redirect)
                Begin
                        Get new URLlocation
                        NumberHops ++;
                        If (NumberHops >5)
                                Begin
                                        Block the webpage
                                End
                        End if
                End
          Go to webpage
        End loop
End
```

Figure 3 Algorithm to block the webpages with more than 5 redirects

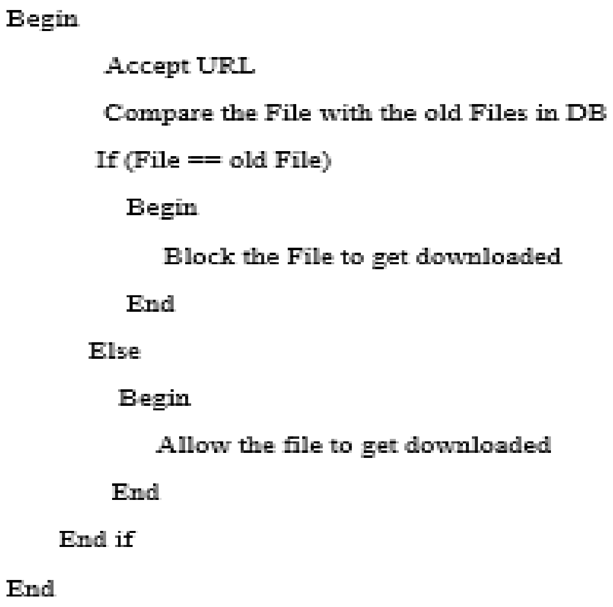
The next algorithm states that when a webpage is entered in the title bar it checks the clusters in the database that the given webpage belongs to the list or not. If not the webpage is redirected to the destination. The algorithm is given below.



```
Begin
    Accept URL
    Compare the URL with the oldURL in DB
    If (URL == oldURL)
        Begin
            Block the user
        End
    Else
        Begin
            Allow the user
        End
    End if
End
```

Figure 4 Algorithm to detect Phishing websites

The following algorithm is for malware samples same as the Phishing websites and it detects with the tag file with the file name. The algorithm is shown below.



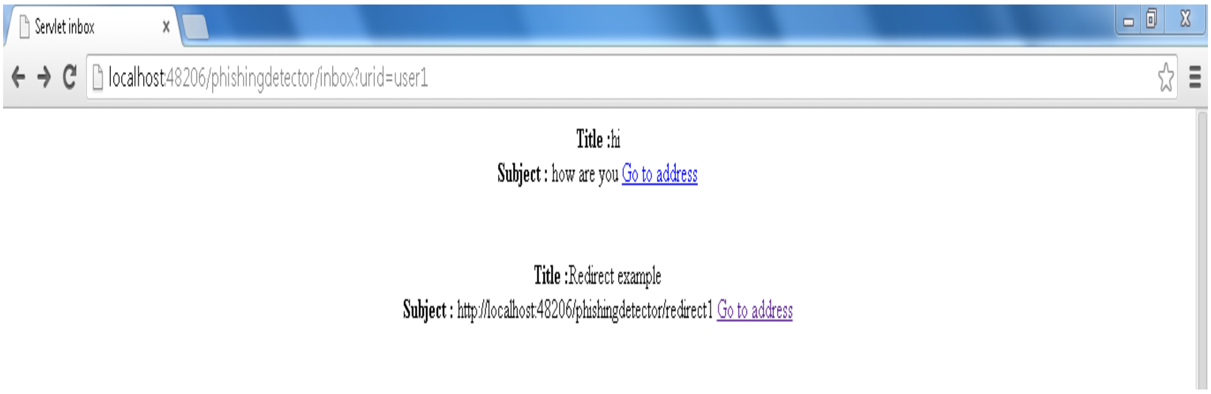
```
Begin
    Accept URL
    Compare the File with the old Files in DB
    If (File == old File)
        Begin
            Block the File to get downloaded
        End
    Else
        Begin
            Allow the file to get downloaded
        End
    End if
End
```

Figure 5 Algorithm to detect Malware samples

4. EXPERIMENTAL RESULTS AND ANALYSIS

The Experimental results of the proposed system are shown. The results are provided with the implementation of clustering techniques BIRCH and DB scan. The Clustering techniques is applied as BIRCH for Malware samples and DB scan for Phishing websites and these are checked in the database where malware samples and Phishing websites are stored and the result is provided.

The first result is the output of the redirects. Here the number of redirects are calculated for the webpages and if the webpage doesn't fit the mark of the given number of redirects then the web page would be blocked and a message will be displayed saying that the email link contains more than given redirects and so it is blocked.



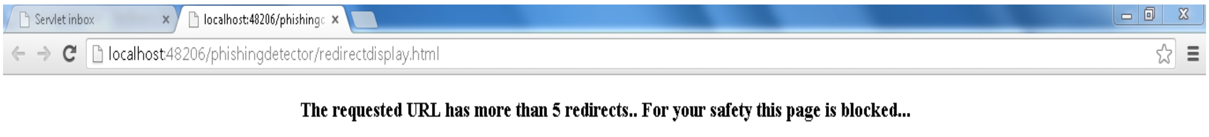
Title :hi

Subject : how are you Go to address

Title : Redirect example

Subject : http://localhost:48206/phishingdetector/redirect1 Go to address

Figure 6: Email with a link attached to it


Servlet inbox
localhost:48206/phishingd
localhost:48206/phishingdetector/redirectdisplay.html

The requested URL has more than 5 redirects.. For your safety this page is blocked...

Figure 7: Email with a link is blocked

The redirection technique is shown above as the email link is blocked if it contains more than 5 redirects. As this number can be extended as much as possible. By reports, the legitimate links would take very small amount of redirects while the non-legitimate users will be using more redirects.

The phishing websites are very dangerous in the internet world as they can appear as a legitimate website and get the details from the user. The result shown below blocks the phishing website with the help of DB scan Technique.

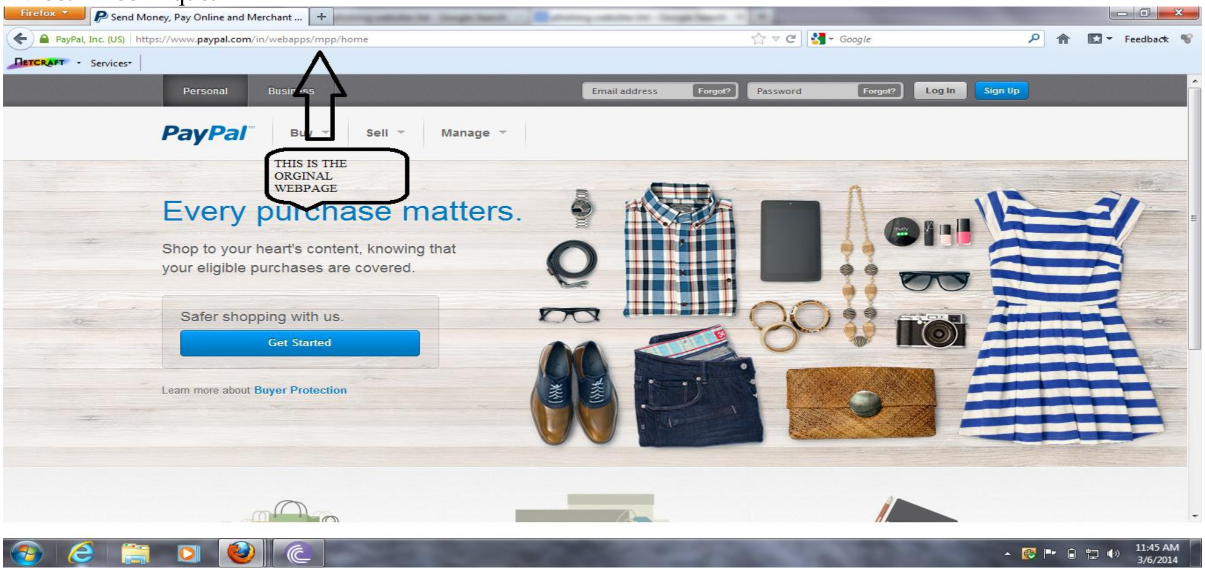

Firefox
Send Money, Pay Online and Merchant ...
PayPal, Inc. (US)
https://www.paypal.com/in/webapps/mpp/home
Google
Feedback
NETCRAFT
Services
Personal
Business
Email address
Forgot?
Password
Forgot?
Log In
Sign Up
PayPal
Buy
Sell
Manage
THIS IS THE ORGINAL WEBPAGE
Every purchase matters.
Shop to your heart's content, knowing that your eligible purchases are covered.
Safer shopping with us.
Get Started
Learn more about Buyer Protection
11:45 AM
3/6/2014

Figure 8: The original webpage of the blocked one

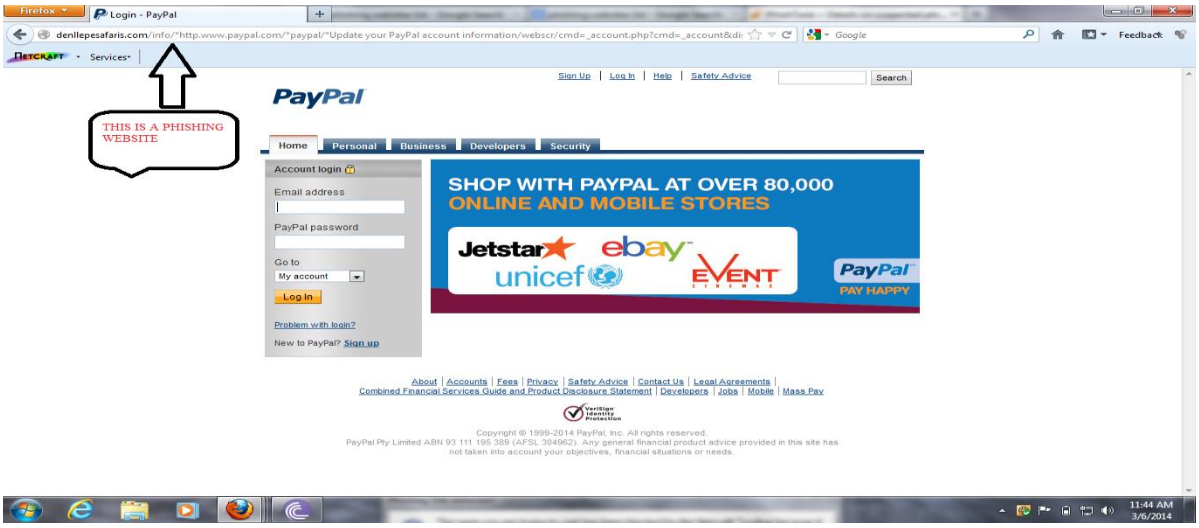
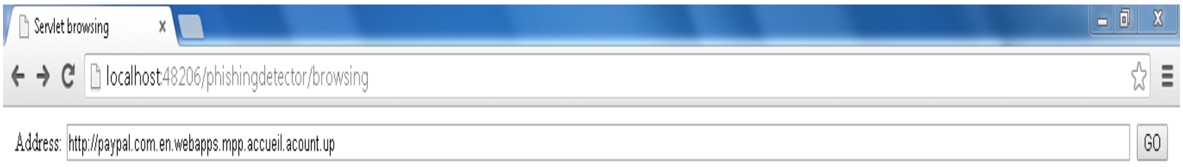

Firefox
Login - PayPal
denllepesafaris.com/info/*http.www.paypal.com/*paypal/*Update your PayPal account information/webscr/cmd=_account.php?cmd=_account&di:
Google
Feedback
NETCRAFT
Services
THIS IS A PHISHING WEBSITE
Sign Up | Log In | Help | Safety Advice
Search
PayPal
Home
Personal
Business
Developers
Security
Account login
Email address
PayPal password
Go to
My account
Log In
Problem with login?
New to PayPal? Sign up
SHOP WITH PAYPAL AT OVER 80,000 ONLINE AND MOBILE STORES
Jetstar
ebay
unicef
EVENT
PayPal
PAY HAPPY
About | Accounts | Fees | Privacy | Safety Advice | Contact Us | Legal Agreements |
Combined Financial Services Guide and Product Disclosure Statement | Developers | Jobs | Mobile | Mass Pay
VeriSign Identity Protection
Copyright © 1999-2014 PayPal, Inc. All rights reserved.
PayPal Pty Limited ABN 93 111 195 389 (AFSL 304962). Any general financial product advice provided in this site has not taken into account your objectives, financial situations or needs.
11:44 AM
3/6/2014

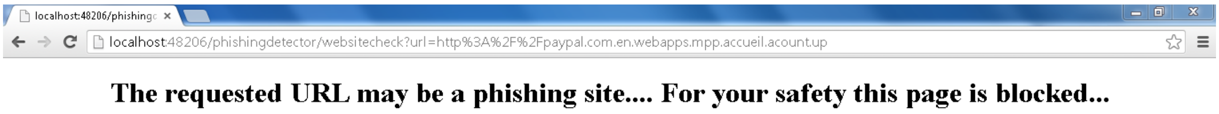
Figure 9: The phishing webpage



localhost:48206/phishingdetector/browsing

Address: http://paypal.com.en.webapps.mpp.accueil.acount.up

Figure 10: A phishing website is entered in the address bar

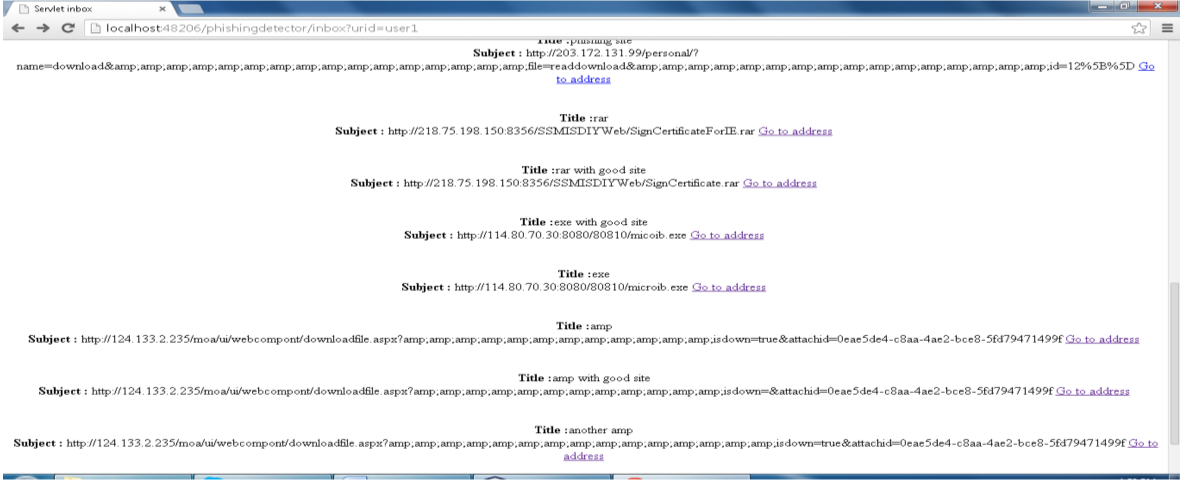


localhost:48206/phishingdetector/websitecheck?url=http%3A%2F%2Fpaypal.com.en.webapps.mpp.accueil.acount.up

The requested URL may be a phishing site.... For your safety this page is blocked...

Figure 11: Phishing website is been blocked

The Malware samples are blocked with the help of the tag name they have with their file name and also these are integrated with the database and the files can be found if it has the same tag name. This is done with the help of BIRCH Clustering.



localhost:48206/phishingdetector/inbox?urid=user1

Title :phishing site
Subject : http://203.172.131.99/personal/?name=download&amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;file=readdownload&amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;id=12%5B%5D Go to address

Title :rar
Subject : http://218.75.198.150:8356/SSMISDIYWeb/SignCertificateForIE.rar Go to address

Title :rar with good site
Subject : http://218.75.198.150:8356/SSMISDIYWeb/SignCertificate.rar Go to address

Title :exe with good site
Subject : http://114.80.70.30:8080/80810/micoib.exe Go to address

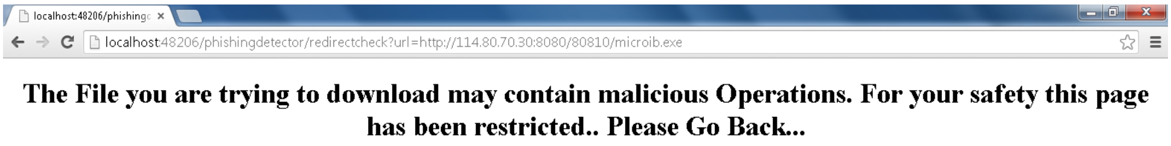
Title :exe
Subject : http://114.80.70.30:8080/80810/microib.exe Go to address

Title :amp
Subject : http://124.133.2.235/moa/ui/webcompont/downloadfile.aspx?amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;isdown=true&attachid=0eae5de4-c8aa-4ae2-bce8-5fd79471499f Go to address

Title :amp with good site
Subject : http://124.133.2.235/moa/ui/webcompont/downloadfile.aspx?amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;isdown=&attachid=0eae5de4-c8aa-4ae2-bce8-5fd79471499f Go to address

Title :another amp
Subject : http://124.133.2.235/moa/ui/webcompont/downloadfile.aspx?amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;isdown=true&attachid=0eae5de4-c8aa-4ae2-bce8-5fd79471499f Go to address

Figure 12: Malware Samples in the inbox



localhost:48206/phishingdetector/redirectcheck?url=http://114.80.70.30:8080/80810/microib.exe

The File you are trying to download may contain malicious Operations. For your safety this page has been restricted.. Please Go Back...

Figure 13: Malware content has been blocked

In the above screen shot, it is clear that the malware content are blocked. The content having the tag name of .exe and .amp amp are checked before it is downloaded. If the content has malware samples then this files are restricted from downloading.

5. CONCLUSION AND FUTURE WORK

There are many antivirus software's in the current software world which gives phishing website detection and malware protection like Kaspersky antivirus, Norton, avast, AVG and Bull guard etc. Netcraft is one add-on which can be installed with the browser and detects the phishing websites. But the ACS described detects more phishing websites and malware samples than the software's and tools listed above. The Automatic categorization system designed can be supported for larger databases which will be useful for the current situation.

With these techniques placed, the users also should have some intelligence about what webpage they are using and what they are clinking on the internet. They should at least know not to provide their SSN number, credit card details, Bank account number and passwords. The users also need to understand the risk if they were providing the passwords and other private matters to others or in unknown sites. In India, Phishing and Malware Spreading is cognizable, bailable and compoundable with permission of the court before which the prosecution of such offence is pending and triable by any magistrate under Information Technology (Amendment) Act, 2008. In the future, the basic clustering techniques can be changed and it can be tested and also the feature extraction techniques can be changed and tested.

REFERENCES

[1] Weiwei Zhuang, Yanfang Ye, Yong Chen, and Tao Li. (2012), 'Ensemble Clustering for internet security applications' in IEEE transactions on systems, man, and cybernetics- part c: applications and reviews, vol. 42, no. 6, November 2012., pp. 1784-1796.

[2] Abu-Nimeh S., Nappa D., Wang X., and Nair S. (2007), 'A comparison of machine learning techniques for phishing detection' in Proc. APWG eCrime Res. Summit, pp. 60–69.

[3] Aburrous M., Hossain M.A., Dahal K. and Thabtah K. (2010), 'Predicting phishing websites using classification mining techniques with experimental case studies' in Proc. 7th Int. Conf. Inf. Technol., pp. 176–181.

[4] Aburrous M., Hossain M.A., Dahal K. and Thabtah K. (2010), 'Predicting phishing websites using classification mining techniques with experimental case studies' in Proc. 7th Int. Conf. Inf. Technol., pp. 176–181.

[5] Azimi J. and Fern X. (2009), 'Adaptive cluster ensemble selection,' in Proc. 21st Int. Joint Conf. Artificial Intelligence. San Francisco, CA, pp. 992–997.

[6] Bailey M., Oberheide J., Andersen J., Mao Z.M., Jahanian F., and Nazario J. (2007), 'Automated classification and analysis of internet malware' in Recent Advances in Intrusion Detection, Vol. 4637, pp. 178–197.

[7] Bayer U., Comparetti P.M., Hlauschek C., Kruege. C., and Kirda E. (2009), 'Scalable, 861ehaviour-based malware clustering' in Proc. 16th Annual Network Distributed Security Symposium

[8] Chou N., Ledesma R., Teraguchi Y., Boneh D., and Mitchell J.C (2004), 'Client side defense against web-based identity theft' in Proc. 11th Annual Network Distrib. Systems Security Symposium

[9] Dazeley R., Yearwood R.L., Kang B.H., and Kelarev A. V (2010), 'Consensus clustering and supervised classification for profiling phishing emails in internet commerce security' in Knowledge Management and Acquisition for Smart Systems and Service, Vol. 6232, pp. 235–246.

[10] http://www.google.com

[11] http://www.wikipedia.com

ABOUT THE AUTHORS

Shabin Blesson was born in 1988. He is currently a student of Francis Xavier College of engineering and technology, Tirunelveli. He is doing his masters in Network engineering. He received his bachelors in electronics and communications from Rajiv Gandhi College of engineering and technology in 2009. His area of interest is network security and his subject of interest is cryptography and network security.

R. Ravi is an Editor in International Journal of Security and its Applications (South Korea). He is presently working as a Professor & Head and Research Centre Head, Department of Computer Science and Engineering, Francis Xavier Engineering College, Tirunelveli. He completed his B.E in Computer Science and Engineering from Thiagarajar College of engineering, Madurai in the year 1994 and M.E in Computer Science and Engineering from Jadavpur Government research University, Kolkatta. He has completed his Ph.D in Networks from Anna University Chennai. He has 18 years of experience in teaching as Professor and Head of department in various colleges. He published 12 International Journals, 1 National Journal. He is also a full time recognized guide for various Universities. Currently he is guiding 18 research scholars. His areas of interest are Virtual Private networks, Networks, Natural Language Processing and Cyber security.

Dr. Beulah Shekhar is a Coordinator for Victimology & Victim Assistance, in the Department of Criminology and Criminal Justice Sciences; she is presently working as an Associate Professor in the Department of Criminology and Criminal Justice Sciences. And her areas of interest are Crimes against Women Empowerment, Human Rights, and Police Training.