EVALUATION OF PHISHING DETECTION ATTACKS WITH URL COMPOSITION

Ravindra Patel
M. Tech Dayscholar
SSSIST Sehore

Mr.Kailash Patidar
HOD CSE Dpartment
SSSIST Sehore

Harsh Lohiya
Assistant Professor
SSSIST Sehore

Abstract:- *Rapid increase in the size of web users. Users enter sensitive information such as passwords, their personal and professional information into scam web sites. Phishing is the criminally fraudulent process of attempting to acquire sensitive information such as usernames, passwords and credit card details, for some illegitimate purpose. Such scam sites cause substantial damages to individuals and corporations. These attacks can be analyzed through this work, and a plug in is designed which provide security from the fake websites . This work is improved by using decision tree c4.5 over id3 and a comparison is drawn.*

I. INTRODUCTION

1.1 What is a Phishing Attack?

While the Internet has brought unprecedented convenience to many people for managing their finances and investments, it also provides opportunities for conducting fraud on a massive scale with little cost to the fraudsters. Fraudsters can manipulate users instead of hardware/software systems, where barriers to technological compromise have increased significantly. Phishing is one of the most widely practised Internet frauds.

It focuses on the theft of sensitive personal information such as passwords and credit card details. Phishing attacks take two forms:

• attempts to deceive victims to cause them to reveal their secrets by pretending to be trustworthy entities with a real need for such information;

• attempts to obtain secrets by planting malware onto victims' machines.

Phishing attacks that proceed by deceiving users are the research focus of this thesis and the term 'phishing attack' will be used to refer to this type of attack.

The number of reported phishing web sites increased 50 percent from January 2008 to January 2010. During the 2008 world financial crisis phishing attack incidents increased three times compared to the same period in 2007.

There is no doubt phishing can be extremely damaging all organizations since tricking a user within a business network through a phishing scam is an easy way to obtain the user's information in order to gain access to that business network. According to the RSA 2012 annual fraud report, the total number of phishing attacks in 2012 was 59% higher than 2011 (RSA, 2012). Global losses from phishing were estimated at $1.5 billion in 2012. That amount of damage is a 22% increase from 2011. The report estimated losses from phishing in 2013 would exceed $2 billion. The following graph in the figure 1.1 shows the number of phishing attacks per year.

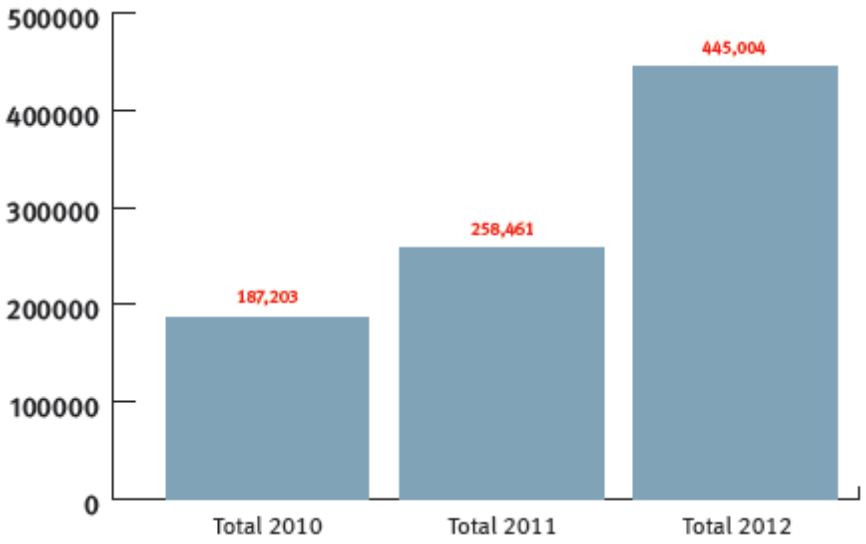

Figure 1 Phishing Attacks per Year

Phishing can also have a large impact on individual Internet users. According to the APWG report, among the top-level domains (TLDs) the .COM namespace contained the most unique domain names used for phishing as well as having the highest number of attacks within the namespace in the quarter of year 2013 (APWG, 2013). This would suggest that a large number of phishing attacks targeted typical Internet users and not corporations. This conclusion is particularly harmful, as typical Internet users have many user accounts on various websites that could be exploited, including accounts for banking, social media, and email. Imperva, a data security company, suggests that users use different passwords for each Internet website that they frequent in order to prevent multiple sets of credentials from being compromised in an attack (Imperva, 2010).

II. Types OF Phishing Attacks

A. Basic URL Obfuscation:URL obfuscation misleads the victims into thinking that a link and/or web site displayed in their web browser or HTML-capable email client is that of a trusted site. These techniques tend to be technically simple yet highly operational, and are still used to some extent in phishing emails today. Some most frequently used methods of phishing attacks are:

- Simple HTML redirection
- Use of JPEG images
- Use of alternate encoding schemes
- Registration of similar domain names

B. Web Browser Spoofing Vulnerabilities:Over the last few years, a number of vulnerabilities in web browsers have provided phishers with the ability to obfuscate URLs

and/or install malware on victim machines. All the vulnerabilities listed currently have fixes available from their associated vendors. However, these vulnerabilities can still be exploited on computers that are not up to date with security patches.

C. International Domain Names (IDN) Abuse:International Domain Names in Applications (IDNA) are a mechanism by which domain names with Unicode characters can be supported in the ASCII format used by the existing DNS infrastructure. A web browser that supports IDNA would interpret this syntax to display the Unicode characters when applicable. Users of web browsers that support IDNA could be susceptible to phishing via homograph attacks;somewhere an attacker could register a domain that contains a Unicode character that appears identical to an ASCII character in a legitimate site. While a proof-of-concept of this type of attack was made public, there has not been any publicly reported IDNA abuse within a phishing scam[4].

D. Web Browser Cross-Zone Vulnerabilities: Most web browsers implement the concept of security regions, where the security settings of a web browser can vary based on the location of the web page being viewed. Author [4] have observed phishing emails that attempt to lure users to a website attempting to install spyware and/or malware onto the victim's computer or device. These web sites usually rely on vulnerabilities in web browsers to install and execute programs on a victim's computer, also when these sites are located in a security zone that is not trusted and normally would not allow those actions.

E. Session Hijacking: Most phishing scams rely on deceiving a user into visiting a malicious web site. However, there is the threat of a user being redirected into a phishing site even if they correctly try to access a legitimate site. Some most frequently used techniques are listed below:

(1) Domain Name Resolving Attacks

(2) Cross-Site Scripting Attacks

(3) Domain Name Typos

(4) Man-in-the-Middle Attacks

F. Abuse of Domain Name Service: Criminals often take advantage dynamic DNS providers, which are often used for providing a static domain name mapping to a dynamic IP address. This service can be useful to phishers by providing them with the ability to easily redirect traffic from one phishing site to another if the initial site is shut down. With ISPs and law enforcement becoming more proactive in shutting down phishing sites, use of dynamic DNS and registration of multiple IP addresses for a single fully qualified domain name (FQDN) is becoming more prevalent to increase the resilience of phishing sites.

G. Specialized Malware: Over the last few years, there has been an emergence of malware being used for criminal activity against users of online banking and commerce sites. This type of specialized malware (which can be considered a class of spyware) greatly increases the potential return on investment for criminals, in case them with the ability to

target information for as many or as few sites as they wish. One advantage for criminals is that most malware can easily be reconfigured to change targeted sites and add new ones. The malware also provides several mechanisms for stealing data that improve the potential for successfully compromising sensitive information. Typical similar kinds of attacks are:

- Electronic Surveillance
- Password Harvesters
- Autonomous Scam Pages and Dialog Boxes
- Account siphons

III. Detecting Phishing Attacks

Identifying phishing attempts is a difficult and unsolved problem due to the inherent vulnerability residing at the receiving end of phishing emails—a human. The prevalence of phishing web sites and emails attests to the success phishers are having with their attempts. When a web site or email emulates a known legitimate site or email, it is relatively easy to fool most Internet users. While phishing training may help the human only slightly, significant advancements are made toward effective technical solutions that are categorized into two groups: content-based filtering and application based filtering.

A. Content-Based Filtering

Content-based filtering refers to statistical analysis, data mining, feature set selection, machine learning, and/or heuristics-based detection mechanisms applied to either email content or web site content.

Fette et al. establish a machine learning algorithm on a feature set designed to highlight human-targeted deception behaviors in email [FST07]. Their approach is named PILFER, and it is a machine learning-based approach to classifying phishing attempts. PILFER uses data directly present in email as well as data collected from external sources. This combined approach creates a feature vector, which is used to train a model for classification. Their feature vector consists of 10 features: Internet Protocol (IP) addresses within web links, age of linked-to domains, non-matching links, ―Here‖ links to a non-modal domain (anomalous links to the non-dominate domain present in the email), HTML (Hyper Text Markup Language) emails, number of links, number of domains, number of dots (e.g., www.this.is.a.bad.site.com), contains JavaScript, and output from third party spam filters. PILFER inputs this feature vector into a random forest as a classifier, where numerous decision tress are created. Preliminary experiments show a 96% detection rate with only a 0.1% false positive rate over 860 phishing and 6,950 non-phishing emails.

L'Huillier et al. propose an online phishing classification scheme using adversarial data mining and signaling games in [LWF09]. They implement a gametheoretic data mining framework that uses dynamic games of incomplete information to build a classifier to detect

phishing attempts. The feature set consists of email content based features, of which there are four categories: email structures related to different email formats, properties of every link in a message, HTML/JavaScript/forms used, and the Spam Assasin's output score for the email in question. This work achieves a high detection accuracy of 99%.

Bergholz et al. propose a number of novel features that are tailored to phishing email detection [BDG+10]. These new features extend the work of L'Hullier et al. by adding a word list to their basic feature set, and advanced graphical features are added as well. These graphical features are image distortion (i.e., attempts to defeat character recognition tools), logo detection (i.e., compared to original logo), and hidden text salting. Hidden text salting consists of random strings, spacing, coloring, spelling, etc. to fool automated appliances but remain invisible to humans. These features are passed into a text classification-based classifier (e.g., random forests or support vector machines). Experiments with these novel features yield a 99.46% accuracy rate, which is slightly higher than that reported by L'Hullier et al.

B. Application-Based Filtering

Application-based filtering refers to a specific method of implementing a phishing detection or prevention mechanism. This category encompasses email client or web browser plugins as well as modified email architecture.

Zhang et al. developed an automated test bed for evaluating anti-phishing tool . They evaluate 10 popular appliance-based anti-phishing tools using 200 phishing URLs (Uniform Resource Locators, or links) from two sources and over 500 legitimate URLs. The results of their evaluation show that only one of the tools could consistently identify over 90% of phishing URLs. However, this same tool also had a 42% false positive rate. In addition, the authors point out numerous methods to exploit vulnerabilities in multiple anti-phishing tools that resulted in phishing sites being labeled as legitimate. Most of the tools use a blacklist of URLs that they would obtain dynamically and frequently. Only one tool uses heuristics-based detection instead of an explicit blacklist. This tool also has high false positive rates. The major contribution of this effort is the authors' conclusion that the success of anti-phishing tools using blacklists relies on very large amounts of data being collected frequently.

Crain et al. propose a tool to assist users in identifying legitimate emails [COP10].

This tool, called Trusted Email, allows companies to establish keys with their clients/customers. This key is used to sign and encrypt emails between the legitimate company and its user. This approach's strength is that it uses existing technology in a novel way to dramatically improve email security. A client-based plugin provides feedback to users when: 1) a key establishment email arrives, 2) a signed email arrives, and 3) a forged email is detected. A small pilot study shows that all users reject all emails marked as phishing, and they also accept all emails that are signed. However, most of them also rejected all unsigned, legitimate emails, which may be a result of the small group of people and their insight into this research.

C. Limitations of Phishing Detection

The content filtering techniques focus their detection on anomalous behavior indicative of phishing. This implies that all phishing attempts use non-standard behavior. However, spear phishing specifically emulates a valid user behaving in a legitimate manner and emailing appropriate recipients. Therefore, the content-based filtering algorithms likely will not recognize legitimate-looking spear phishing emails.

On the other hand, application-based filtering shows promise, but it relies heavily on the use of blacklists that must be constantly updated. But there is an inherent challenge with this: any new phishing attempts will have to be discovered first before it can be added to a known bad blacklist. Even heuristic-based detection suffers from unacceptable false positive rates. Therefore, current application-based and content filtering-based phishing detection techniques likely will not catch spear phishing attacks, especially ones crafted and targeted for the purpose of cyber espionage.

IV. Existing work

There are various techniques are available to detect and prevent the phishing attacks some efforts in this domain is listed below.

machine learning technique:An study to compares the predictive accuracy of several machine learning methods including Logistic Regression (LR), Classification and Regression Trees(CART), Bayesian Additive Regression Trees (BART), Support Vector Machines (SVM), Random Forests (RF), andNeural Networks (NNet) for predicting phishing emails. For evaluation of given model a data set of 2889 phishing and legitimate emails are used inthe comparative study. In addition, 43 structures are used to train and test the classifiers. The application of these algorithms is to classify the phishing attack emails is found.

Content based approach:In this type of approaches content of web pages or structure of web pages are extracted and compared, present the design, implementation, and evaluation of CANTINA framework, which is a novel, content-based approach to detecting phishing web sites, that is based on the TF-IDF information retrieval algorithm. Where author also discuss the design and evaluation of several heuristics they developed to reduce false positives. Given experiments show that CANTINA is good at detecting phishing sites, correctly labelling approximately 95% of phishing sites.

Image based verification schemes: In this technique an image is used to cross verify the authenticity of user that allow a human to distinguish one computer fromanother. Different traditional HIPs, where the computer concerns a challengeto the user over a network, in this case, the user concerns a challenge to thecomputer. This category of HIP can be used to detect phishing attacks, where websites are spoofed in order to trick users into revealing private information. A new anti-phishing proposal, provides Dynamic Security Skins (DSS), and show that it meets the HIP criteria. Authors goal is to allow a remoteserver to prove its identity in a way that is easy for a human user toverify and hard for an attacker to spoof. In this scheme, the web serverpresents its proof in the form of an image that is unique for each

userand each transaction. For authentication of the server, user can visuallyverify that the image presented by the server matches a reference imagepresented by the browser.

In addition of that more than 100 techniques are found for detection and prevention of phishing attack.

V. System architecture

To provide the optimum solution for the Anti-phishing we proposed the below given system architecture. To properly understand and implement the complete model some modules are designed, their description and working is given as:

Phish tank Database: that is an updated database where the entire phish reported web URLs is stored, proposed system contains a relational data table which store these web URL patterns and used to build data model form algorithm selected.

Universal Database: this database in common for all guests who use proposed tool, this database contains user feedback about URLs.

Navigated URL: that is a user interface where user navigated URL information is stored and provides the various user experiences about the navigated page.

USER Feedback: a user interface provided in the proposed model to submit feedback for a URL if required to report and this is taken in both databases.

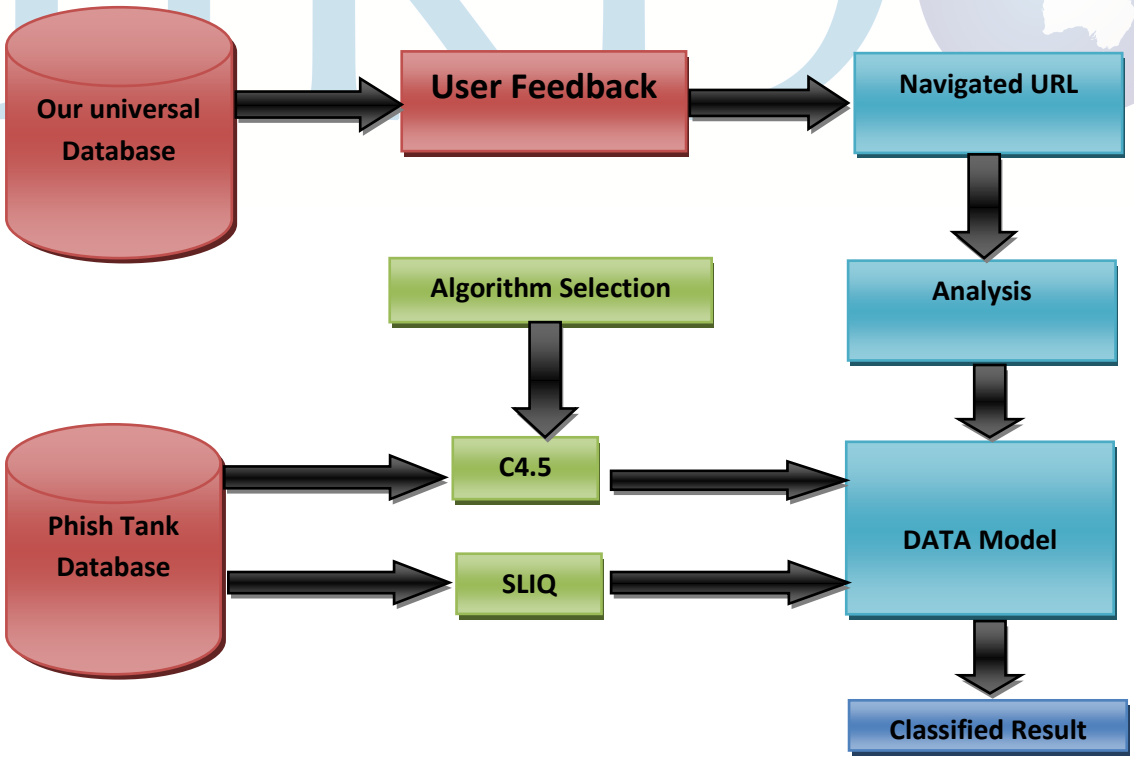

Fig 2System architecture

Algorithm selection: this module contains algorithms and user select an algorithm for consuming phish tank database and develop a data model for navigation.

Data model: the developed data model is a decision tree which is grown using a phish tank database and used to analysis the URL pattern which is found in the database. After analysis of web URL decision is reached.

VI. Results Analysis

The given section of the document includes the performance of the classifiers that are implemented in the current anti-phishing browser extension. Various performance parameters that are required to evaluate for performance analysis, provided results are the classification performance due to continuously increasing data.

1. Classification Accuracy

The calculated performance of the proposed system in terms of accuracy is measured in terms of percentage which is evaluated using n cross validation method. The overall classification accuracy is given below and evaluated using the below given formula, the listed accuracy of the system are the best performance during different experiments.

$$\%\ \text{accurecy} = \frac{Total\ correctly\ classified}{total\ values\ to\ classify} X100$$

The comparative results show the performance graph using figure 6.1 which demonstrates that the performance of the system is most of the time system performance is not varying in large quantity and keep studying in all conditions.

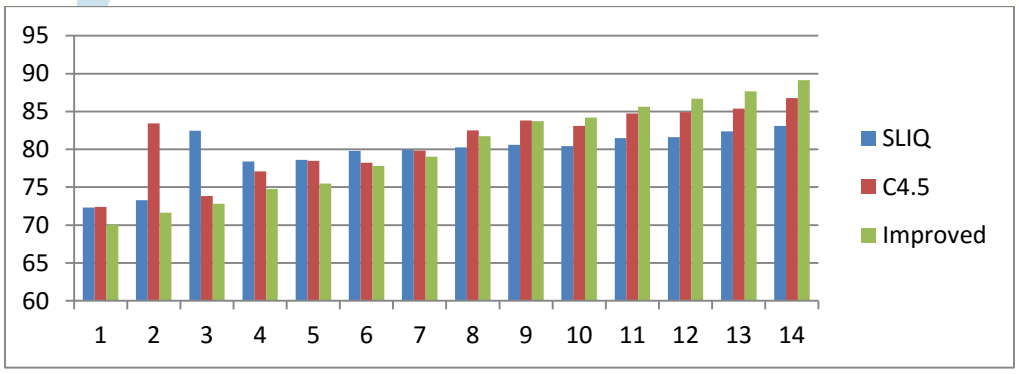

Fig 5.1 shows the accuracy of the system

5.4.3 Memory Uses

That is defined as the memory resources consumed during the performance of the system, here the consumption of the resources in giving in KB.

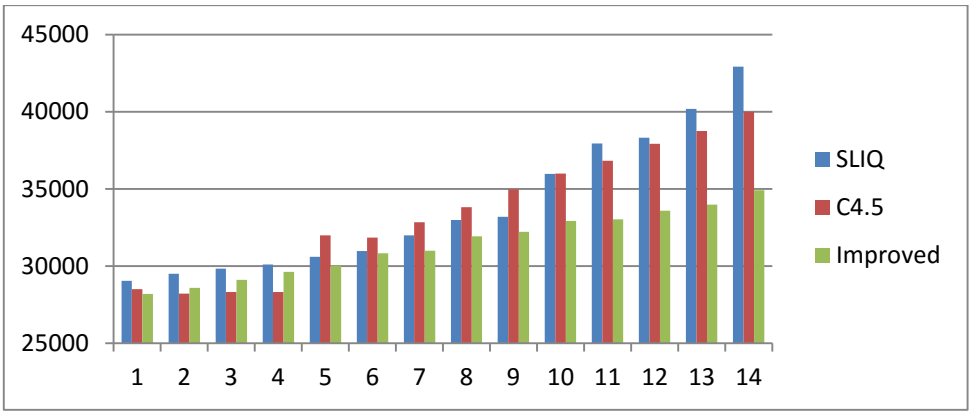

Fig 5.3 shows the memory consumed

Memory consumption of both classifiers are consumes about similar performance and increasing as the size of data in main memory is increases.

6.2 Conclusion and Future Work

The proposed study work provides the efforts for search an efficient and effective anti-phishing tool. This work includes the literature study and collection for finding the attacks and their characteristics, after concluding various research papers we found that the prevention is depends upon the kind and type of attack. Thus to detect the URL based phishing attacks and abuse propose a new solution which is based on a global phishing database (obtained from phish tank database), browser extension and indicator, user feedback and data mining based hybrid approach the proposed system is promises to detect and prevent session high jacking kinds of attack.

Proposed Anti-phishing tool is implemented successfully and working as expected at the time of design. This tool includes all the aspects which is proposed in this document, additionally the proposed method is efficient and providing much accurate results with low memory and time resource consumption. All software tools contains some bugs and modifications are arises due to platform and hardware changes thus this system also need some time required modifications. Additionally required to collect more literature collection according to the time change domain of phishing types and their prevention schemes, by with system needs to enhance as new kind of attack is found in the current web systems.

REFRENCES

[1] Amir Herzberg and Ahmad Jbara, “Security and Identification Indicators for Browsers against Spoofing and Phishing Attacks”,Security and Identification Indicators – Version of 9/3/2006

[2]PonnurangamKumaraguru, Justin Cranshaw, Alessandro Acquisti, Lorrie Cranor, Jason Hong, Mary Ann Blair, Theodore Pham, "School of Phish: A Real-Word Evaluation of Anti-Phishing Training",March 9, 2009, CMU-CyLab-09-002, CyLab- Carnegie Mellon University Pittsburgh, PA 15213

[3]SergeEgelman, Lorrie Faith Cranor, Jason Hong,"You've Been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings",Copyright 2008 ACM 1-59593-178-3/07/0004

[4] Jason Milletary, Technical Trends in Phishing Attacks,US-CERT,https://www.us-cert.gov/reading_room/phishing_trends0511.pdf

[5] Joshua S. White, Jeanna N. Matthews, John L. Stacy ,"A Method For The Automated Detection Of Phishing Websites Through Both Site Characteristics And Image Analysis",http://people.clarkson.edu/~jmatthew/publications/SPIE_PhishingDetection_2012.pdf

[6] Saeed Abu-Nimeh, Dario Nappa, Xinlei Wang, and Suku Nair, A Comparison of Machine Learning Techniques for Phishing Detection, October 4-5, 2007, Pitts-burgh, PA, USA.

[7] Maher Aburrous, M. A. Hossain, KeshavDahal, FadiThabtah,Predicting Phishing Websites using Classification Mining Techniques with Experimental Case Studies,http://scim.brad.ac.uk/staff/pdf/mahossa1/ITNG%202010%20Conference%20Paper.pdf

[8]Ian Fette, Norman Sadeh,Anthony Tomasic, "Learning to Detect Phishing Emails",2007, May 8–12, 2007, Banff, Alberta, Canada.ACM 978-1-59593-654-7/07/0005.

[9]Collin Jackson, Daniel R. Simon, Desney S. Tan, and Adam Barth,An Evaluation of Extended Validation andPicture-in-Picture Phishing Attacks,Microsoft Research, Redmond, WA,2007 – Springer,http://usablesecurity.org/papers/jackson.pdf

[10]ChuanYue and Haining Wang, "Anti-Phishing in Offense and Defense",1063-9527/08 $25.00 © 2008 IEEE

[11]Saeed Abu-Nimeh, Dario Nappa, Xinlei Wang, Suku Nair, "A Distributed Architecture for Phishing Detectionusing Bayesian Additive Regression Trees", 978-1-4244-2969-1/08, 2008 IEEE

[12] Maher Aburrous,M. A. Hossain,KeshavDahal,FadiThabatah, "Modelling Intelligent Phishing Detection System for e-Banking using Fuzzy Data Mining", 978-0-7695-3791-7/09 $26.00 © 2009 IEEE

[13]John Yearwood, Musa Mammadov and Arunava Banerjee, "Profiling Phishing Emails Based on HyperlinkInformation",978-0-7695-4138-9/10 $26.00 © 2010

[14]Kris Beck, Justin Zhan, "Phishing Using A Modified Bayesian Technique", 978-0-7695-4211-9/10 $26.00 © 2010 IEEE

[15]Tianyang Li, Fuye Han, Shuai Ding and Zhen Chen, "LARX: Large-scale Anti-phishing by RetrospectiveData-Exploring Based on a Cloud ComputingPlatform", 978-1-4577-0638-7 /11/$26.00 ©2011 IEEE

[16]Eric Lin, Saul Greenberg, Eileah Trotter, David Ma and John Aycock, "Does Domain Highlighting Help People Identify Phishing Sites?",Copyright 2011 ACM 978-1-4503-0267-8/11/05.

[17]Jun Ho Huh and Hyoungshick Kim, "Phishing Detection with Popular SearchEngines: Simple and Effective", Springer-Verlag Berlin Heidelberg 2011

[18]Cao, Lianjie; Probst, Thibaut; and Kompella, Ramana, "PhishLive: A View of Phishing and Malware Attacks from an Edge Router", (2012).Computer Science Technical Reports.Paper 1761.http://docs.lib.purdue.edu/cstech/1761

[19]Cheng Hsin Hsu,Polo Wang,Samuel Pu,"Identify Fixed-Path Phishing Attack by STC",Copyright 2011 ACM 978-1-4503-0788-8/11/09

[20]Syed Imran Ahmed Qadri, Prof.KiranPandey, "Tag Based Client Side Detection of Content Sniffing Attacks with FileEncryption and File Splitter Technique",International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Volume-2 Number-3 Issue-5 September-2012

[21] Ammar Almomani, B. B. Gupta, Samer Atawneh, A. Meulenberg, and Eman Almomani "A Survey of Phishing Email Filtering Techniques"pp2070-2081, IEEE Communications Surveys & Tutorials, Vol. 15, No. 4, Fourth Quarter 2013

[22] Pranal C.Tayade, Prof. Avinash P.Wadhe "Review Paper on Privacy Preservation through Phishing Email Filter" International Journal of Engineering Trends and Technology (IJETT) – Volume 9 Number 12 - Mar 2014

[23] Samanjeet Kaur, Sukhwinder Sharma "Detection of Phishing Websites using the Hybrid Approach" International Journal For Advance Research In Engineering And Technology, ISSN 2320-6802,Volume 3, Issue VIII, Aug 2015

[24] S.S. Kulkarni, Mayank Tomar, Aastha Mittal, Sneha Arondekar, Aniket Nayakawadi "Survey on Phishing Attacks Survey on Phishing Attacks" International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X ,Volume 5, Issue 2, February 2015