AN ENHANCEMENT OF ASSOCIATION CLASSIFICATION ALGORITHM FOR IDENTIFYING PHISHING WEBSITES

G.Parthasarathy[1]
Research Scholar, Dept. of CSE
Sathyabama University, Chennai, India
amburgps@gmail.com

D.C.Tomar[2]
Professor, Dept. of Information Technology
Jerusalem College of Engineering, Chennai, India.
dctomar@gmail.com

K. Christina Praisy[3]
Student, Dept. of CSE
Jeppiaar Maamallan Institute of Technology, Sriperumbudur, India.
christinapraisy23@gmail.com

Abstract - Phishing is a fraudulent activity that involves attacker creating a model of an existing web page in order to get more important information similar to credit card details, passwords etc., of the users. This paper is an enhancement of the existing association classification algorithm to detect the phishing websites. We can enhance the accuracy to a greater extent by applying the association rules into classification. In addition, we can also obtain some valuable information and rules which cannot be captured by using any other classification approaches. However the rule generation procedure is very time consuming while encountering large data set. The proposed algorithm makes use of Apriori algorithm for identifying frequent itemsets and hence derives a decision tree based on the features of URL.

Keywords- Phishing; Classification; web mining;

I. INTRODUCTION

The growth of internet was unprecedented at the time it was invented. It spread like a wild fire all across the globe, in a short span of time. Online services have made our life much easier. We can shop from our house and the product will be delivered at our door steps. We can also make use of online banking systems for making money transactions. Unfortunately even with the advancement of technologies, certain people invade the internet securities and gain unauthorized access to user's private information.

Phishing is a fraudulent activity in which certain people try to gain unauthorized information of the users via internet. The word 'phishing' is a homophone of the word 'fishing' due to the similarity of using bait/fake website in an attempt to capture fish/victim. The people who are involved in such activities are called phishers. The process of phishing involves phishers sending out a bulk of spoofed emails to the users. When the user clicks the hyperlink provided in the mail, the page is redirected to the fake website created by them. The website will be very similar to the original site that the user will not be able to identify it. The web page usually requests user to provide confidential information such as account number. When the user provides the information, the phishers will gather their data and use it for other malicious activities. We can identify the phishing website only with the help of its URL.

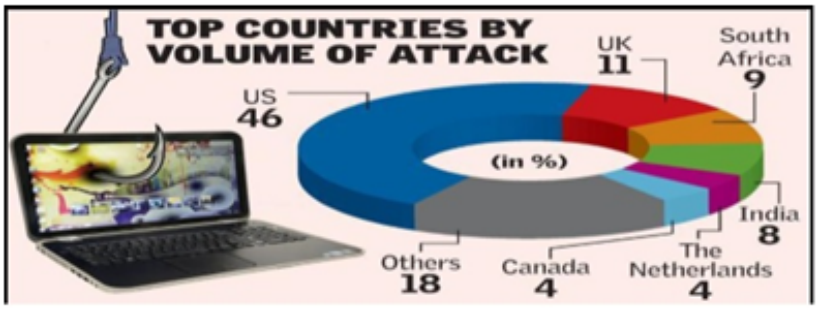

Figure 1 Phishing attack countries ranked.

The above figure 1 depicts the top countries which are under the threat of phishing attacks in terms of volume of attack. The US is ranked first among the countries targeted by phishers. India is ranked fourth in the phishing attacks.

The detection of phishing can be achieved by either increasing user awareness or using software based detection. Although there are several software detection techniques that address the problem of phishing detection, phishing has become more and more complicated and sophisticated, and can bypass the filter set by anti-phishing techniques [14]. Phishtank is a website that holds the list of phishing websites. It contains the list of blacklisted websites which are either active or inactive. An important factor for the effectiveness of a blacklist [15] is its coverage. The coverage specifies how various phishing pages on the Internet are incorporated in the list. A new issue is the excellence of the list. The excellence indicates how several non-phishing sites are incorrectly incorporated into the directory for each incorrect entry; the user experiences a false warning when he/she visits a legal site, undermining his/her trust in the usefulness and correctness of the result.

In this article we target on classifying websites based on the features of URL. We make use of Apriori algorithm to identify frequent itemsets. Here the itemsets are the features of the URL. Then a decision tree is formed which is used for classifying the websites into legitimate website or not. This method simplifies the use of association classification algorithm to derive association rules yet it is more effective than the previous approaches.

II. RELATED WORK

Many anti-phishing techniques have been adopted for efficient detection of phishing websites. Some approaches involve preventing such emails from reaching the victims. It involves anti-spam techniques for filtering and content analysis. Microsoft and Yahoo have also defined e-mail authentication protocols (i.e., Sender ID [1] and DomainKeys [2]) which can be used to verify if a received e-mail is authentic. The main disadvantage of these solutions, however, is that they are currently not used by the majority of Internet users. Another existing approach is Blacklisting. In this method a database with the URLs of phishing websites is maintained. It is a time consuming process since it involves collecting feedback about those sites. Websites such as Phish tank [3] and netcraft [4] contain blacklisted websites. Drawback of this approach is its inability to detect newly created phishing sites.

Client side software such as SpoofGuard and PwdHash are also available. SpoofGuard [5] looks for abnormal URLs in the web pages and raises alert. PwdHash [6, 7] creates domain specific passwords which will be useless when they are submitted to a different domain (e.g. a password for www.gmail.com will be different when submitted to www.phishmail.com). Our approach focuses on using server side programming to detect phishing websites. The advantage that the server side programming has over client side is that by installing single software in the server, thousands of clients can be beneficed. Whereas client side software must be installed in each and every system to offer protection. Hence the latter approach is followed by most of the professionals.

Data mining is the latest approach for detecting phishing sites. In Heuristic-based methods, some features are extracted from the websites which are then classified as either phishy or legitimate, the accuracy depends on features selected. Detailed study on Feature Extraction or Feature Selection for Text Classification to identify a Phishing mail is also an approach for classification. Another approach proposed, utilizes CANTINA (Carnegie Mellon Anti-phishing and Network Analysis Tool) [8] for detection of phishy websites using the concepts of information retrieval measures such as content based techniques. In the existing approach [9], association classification algorithms such as Naive Bayes Classifier, decision tree based J48 algorithm, Meta classifiers AdaBoost and Random Forest were used.

Table 1. A comparison of the existing systems used to detect phishing websites:

Methods	False positives	Zero day attacks	Fake interface attack	Slow response time
Blacklist	No	Yes	No	No
Heuristics	Yes	May be	No	May be
User polling	Yes	Yes	Yes	May be
Third-party certification	No	No	Yes	May be

III. PHASES OF PHISHING

The phishing cycle has 3 main phases [17]. At first, the phisher creates a phishing website similar to the original one. Then he does phishing by transfer the numerous emails to unsuspecting users. The phisher attempts to persuade the reader to visit the web page whose link has been included in the email. While the client "bites" on the phish, the hyperlink in the mail expresses the user to the phishing website which is planned to be same or identical to the legitimate destination site. The phish said to be successful when users enter their confidential information on the phishing page which is then leaked to the phisher. At a later time, the phisher may try to develop the secret data by opening accounts, sending money, or making purchases using the captured data. Sometimes the phisher will just act as a middleman who put up for sale the data to other illegal users.

3.1 THE PHISH

The phishing life cycle begins with a corpus email that tries to induce the reader to monitor the website link incorporated in the email. This phase of phishing is similar to fishing. As a replacement for using a fishing attract and procession to fetch a fish, a phisher throw away several emails in hopes that a small number of readers will "bite" at the email lure by monitoring the link to the phishing website incorporated in the mail. The email is designed to look like a legitimate email and will contain a company logo of an admired financial organization and a return address of the valid company. The website hyperlink in the mail will also seem to be valid at first look. The phisher wants the attract to be as valid as possible so that the victim will "bite". Usually the phishing email will attempt to induce the reader to visit the integrated website in order to modify certain account data or to avoid termination of account. Many such methods are used, but the majority attempt to encourage the reader that imperative action is necessary and kill upon sentiment such as importance, account disruption, or account killing.

3.2THE BITE

The bite occurs when the sufferer clicks on the hyperlink in the mail and depart to the phishing website. The phishing website will look same or much related to the legitimate website which it is attempting to imitate. It is critical for the website to look legitimate so that the users will not believe that the page is unoriginal. frequently the genuine page is basically reproduction and hosted somewhere else so the phishing page having all of the correct styles and information of the genuine website. Logos, keywords, and still protection notices are found on phishing sites to create the client accept as true that the site is genuine. Once the user watches the phishing website and accepts that the page is legitimate since it resembles the legitimate site, then the phisher can demand personal data. It is a critical step for the phisher to first develop and trust so that the user thinks that the page is legitimate. If there are any misspellings, old-fashioned images, or other doubtful datat then the user may think twice about entering sensitive data.

3.3THE CATCH

Once the user has visited the phishing webpage and gets convinced that the webpage is familiar and genuine, then the phisher will request the top secret data from the user. Often there is a user login and password sector in the webpage that needs a username and password from the user. Occasionally a phishing page will request for other top secret data such as account numbers, pins, social security numbers, date of birth, etc. Once the user divulges this data it is stored in a database on the phishing server, or emailed to the phisher's email address, or transfer to a chat room. After the data is submitted the user will obtain an fault message, return to the phishing login box with the impression that not anything happened, or be redirected to the legitimate website. It will show to the user that not anything has happened even though some data has been escaped. Obviously the phisher doesn't would like the user to identify that they have just divulged their secrete data. Phishers collect the secrete data and then either attempt to exploit the data by sending funds, making purchases, etc. or they trade the data to third party frauds to exploit [2]. Underground internet chat rooms are general meeting areas where the phishers

can trade confidential data to parties interested. Phishing sites are live for a very short period before they are discovered and shutdown. Most sites are live for as a few hours to as extended as a few days [12]. Generally the website will be reported and confirmed phishing. Then the ISP will remove the phishing website. But it is hard to delete the website or prosecute the phisher if the website is hosted in a foreign country due to the differing laws and jurisdiction. Normally the financial institutions will reimbursement the lost currency from customers as it is easier and costs less than finding and prosecuting the criminals.

IV. TECHNIQUES OF DETECTING PHISHING WEBSITES

Phishers create an exact replica of the original website. Hence it is not possible to detect phishing websites visually. We make use of some of the important features of URL to make this possible.

TABLE 2. SOME OF THE FEATURES OF THE URL OF THE PHISHING WEBSITE

Serial No.	Behavior	Method
1	With IP address	The main aim of phishers is to gain lot of money with no investment and they will not invest to buy domain names for their fake website. Most phishing websites contain IP address as their domain name. If IP address is used in the domain name, then it is a phishing site.
2	With URL size	Legitimate websites do not have URL of more than 75 characters.
3	Having @ symbol	Websites having @ symbol in its URL are considered to be phishy.
4	Double slash	URLs containing double slash are used for redirecting. Hence they are phishy in nature.
5	Using dash	Legitimate websites do not use dash in its URL.
6	Having sub domain	Legitimate Websites use only domain generally up to two levels. Websites having more than 3 dots do it to include more domains within a domain.
7	Usage of SSL socket layer	Legitimate websites use SSL socket layer every time sensitive information is transferring. Sites not including this can be categorized as Phishing Websites.
8	Port	Phishing Websites make use of some port numbers like 82, and therefore port scanning can help in identifying websites which are phishy.
9	HTTPs token	The HTTPs presence on websites during transfer is a clear measure of authenticity. Phishing websites don't use HTTPs.
10	Request URL	In legitimate websites, objects within same domain are linked to the same domain, once the <src> tag is known, however in Phishy websites; it has been observed that objects are from different domains.
11	Redirect	If the numbers of redirects are more than three, then website can be classified as Phishing Website.
12	Web traffic	If the website is either having no traffic or limited, then it can be classified under Phishing Website.
13	Age of domain	Legitimate websites have an age of six months; websites with more than this age can be classified as Phishy.
14	Page rank	Phishing websites will have low page rank due to lack of links pointing to them.
15	Abnormal URL	This feature is extracted from Whois Database[10] , Legitimate website's main identity is in the URL
16	Nil anchor	Nil anchors denote that the page is linked with none. The value of the href attribute of tag will be null. The values that denote nil anchor are about: blank, javascript:; JavaScript:, void(0),#.
17	Foreign anchor	An anchor tag contains href attribute whose value is a url to which the page is linked with. If the domain name in the url is not similar to the domain in page url then it is called as foreign anchor.

V. PROPOSED SYSTEM

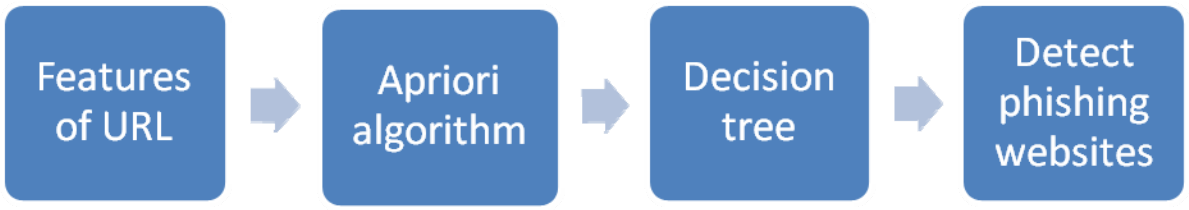

Figure 2. Overview of Detecting Phishing websites

A. Apriori algorithm:

Apriori Algorithm Association rule generation is divided into two steps:

1. First, minimum support is applied to find all frequent itemsets in a database.

2. Second, these frequent itemsets and the minimum confidence constraint are used to form rules.

After finding all frequent itemsets in a database, it becomes very difficult as the process involves searching all possible itemsets (item combinations). The set containing possible itemsets is the power set over I and has size 2n 1 (which excluds the empty set as it is not a valid itemset). Even though the size of the powerset grows exponentially with the number of items n in I, efficient search is possible using the downward-closure property of support (also called anti-monotonicity). It guarantees that for a frequent itemset, all its subsets are frequent. It also guarantees that for an infrequent itemset, all its supersets are also infrequent. Exploiting this particular property, efficient algorithms (e.g., Apriori and Eclat) can find all frequent itemsets. Apriori Algorithm Pseudocode procedure Apriori (T, minSupport) //T is the database and minSupport is the minimum support L1= frequent items; for (k= 2; Lk-1 !=; k++) Ck= candidates generated from Lk-1 //that iscartesian product Lk-1 x Lk-1 and eliminating any k-1 size itemset that is not //frequent for each transaction t in database do #increment the count of all candidates in Ck that are contained in t Lk = candidates in Ck with minSupport //end for each//end for return ; As it is common in association rule mining, given a set of itemsets (for instance, sets of retail transactions, each listing individual items purchased), the algorithm attempts to find subsets which are common to at least a minimum number C of the itemsets. Apriori uses a “bottom up” approach. In this approach frequent subsets are extended one item at a time. This step is known as candidate generation. Then groups of candidates are tested against the data. When no further successful extensions are found the algorithm is terminated.

B. Decision tree algorithm:

The leaf nodes in a decision tree will contain the class name. A non-leaf node must be a decision node. The decision node is an attribute test. Each branch to another decision tree is a possible value of the attribute. ID3 uses information gain to decide which attribute goes into a decision node. The advantage of learning a decision tree is that a program, rather than a knowledge engineer, elicits knowledge from an expert.

ID3 is based off the Concept Learning System (CLS) algorithm. The basic CLS algorithm over a set of training instances C:

Step 1: If all instances in C are positive, then create YES node and halt.

If all instances in C are negative, create a NO node and halt.

Otherwise select a feature, F with values v1, ..., vn and create a decision node.

Step 2: Partition the training instances in C into subsets C1, C2, ..., Cn according to the values of V.

Step 3: apply the algorithm recursively to each of the sets Ci.

Note, the trainer (the expert) decides which feature to select.

ID3 improves on CLS by adding a feature selection heuristic. ID3 searches through the attributes of the training instances and extracts the attribute that best separates the given examples. If the attribute perfectly classifies the training sets then ID3 stops; otherwise it recursively operates on the n (where n = number of possible values of an attribute) partitioned subsets to get their "best" attribute. The algorithm uses a greedy search, that is, it picks the best attribute and never looks back to reconsider earlier choices.

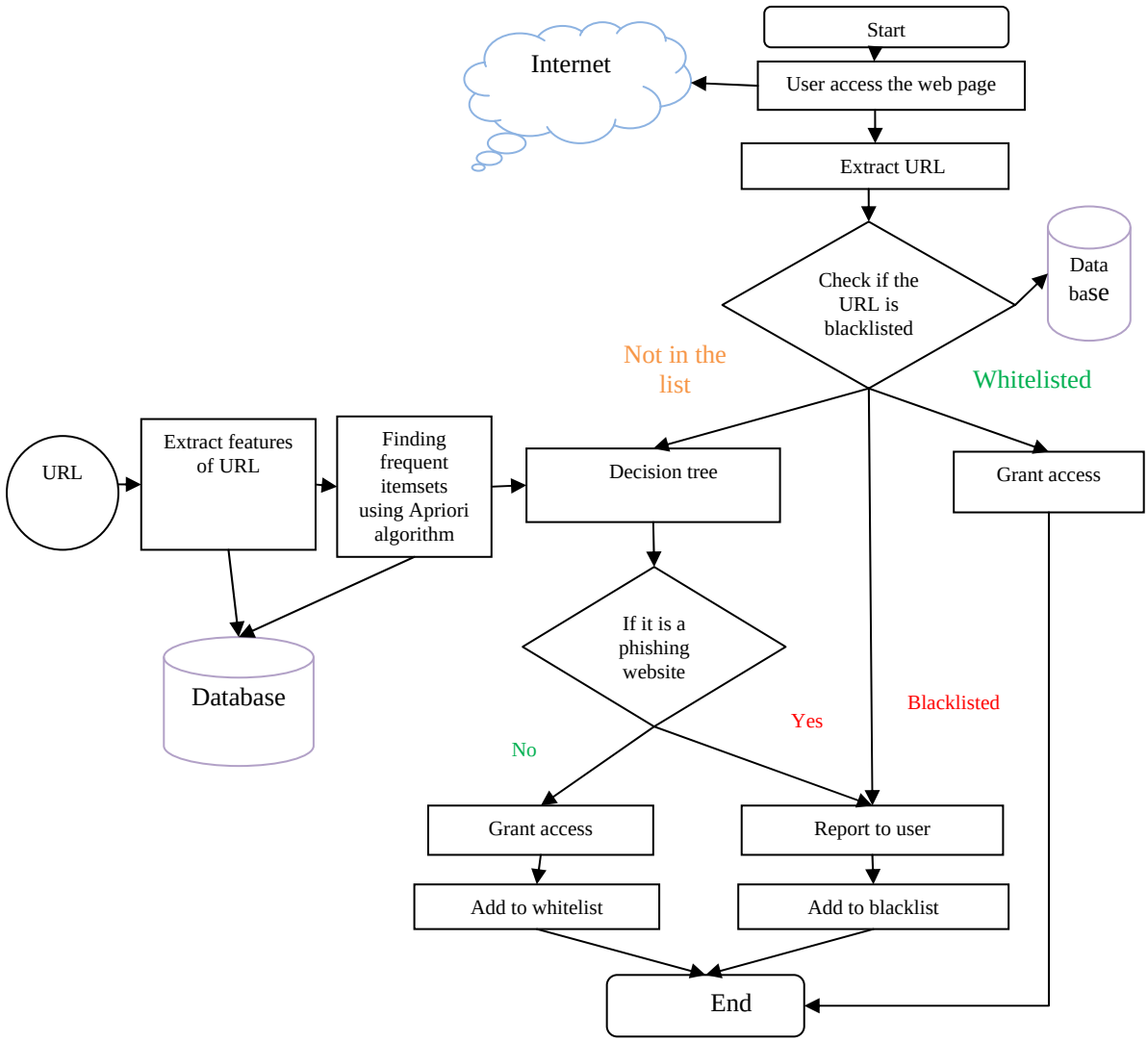

Figure 3 Architecture of Detecting Phishing websites

When user tries to access a webpage, the server extracts the URL and checks if it is blacklisted in the database it maintains. If it is blacklisted, it is reported to the user. And if the site is Whitelisted, access is granted. This will avoid running the proposed algorithm every time user visits the site and hence reduces time complexity. The proposed algorithm focuses only on the websites which are not present in the lists or those sites which are newly created.

Data is preprocessed using the proposed algorithm. For deriving a decision tree, several features of blacklisted URL are collected and ID3 algorithm is applied.

VI. EXPERIMENT AND RESULT

A. Feature extraction:

Features of the url were extracted from various websites such as Phishtank and Netcraft. They were stored in a database for easier access and retrieval. Other features such as url length, presence of @ symbol etc., were calculated using a simple python program.

B. Association algorithm:

Apriori algorithm is used for association classification. It gave important association rules based on the features of url. The algorithm makes use of number of occurrence of the features and minimum support and confidence. The minimum support is based on the probability of a feature having an effect on the url being a phishing website.

C. Decision tree algorithm:

Based on the association rules derived from the application of Apriori algorithm, a decision tree was obtained using ID3 algorithm. The decision tree consists of features related together based on association rules in the nodes and the conditions as links.

D. Detection of phishing sites:

When the user tries to access a website, the server implements the following program. The program involves retrieving the url, followed by extracting the features of url. This program is implemented dynamically during the run time. The decision tree is parsed and it is found out if it is a phishing website or not.

E. Result:

The association classification algorithm and decision tree algorithm were implemented successfully. The result analysis shows that the proposed algorithm was more effective than the previous approaches. This approach can be used dynamically during the run time. Hence servers can install the program in their system and provide a better service for the clients. This approach can be used in the banking servers for gaining trusts of the clients.

CONCLUSION

Classification Data Mining (DM) Techniques are considered to be a highly useful tool in detecting and identifying phishing websites for e-banking sites. In this paper, we are presenting an advance method to conquer the difficulty and complexity involved in finding and forecasting e-banking phishing website. We proposed an bright resilient and successful model that is based on association classification Data Mining methods. These algorithms are used to distinguish and recognize all the factors and rules for the reason of classifying the phishing website and the relationship them with each other. We also compared their performances, accuracy, number of rules generated and speed. A Phishing Case study was applied to demonstrate the website phishing procedure. The rules generated from the associative classification replica showed the relationship between some important characteristics like URL and Domain Identity, and Security and Encryption criteria in the final phishing detection rate. The experimental results shown the feasibility of using Associative Classification techniques in actual applications and its better performance as compared to other conventional classifications.

REFERENCES

[1] Yahoo. Yahoo! AntiSpam Resource Center. http://antispam.yahoo.com/ domainkeys
[2] Microsoft. Sender ID Home Page. http://www.microsoft.com/mscorp/safety/ technologies/senderid/default.mspx
[3] PhishTank Available: https://www.phishtank.com/
[4] Net Craft Available: http://www.netcraft.com/
[5] SpoofGuard. Client-side defense against web-based identity theft. http:// crypto.stanford.edu/SpoofGuard/ 2005.
[6] Blake Ross, Collin Jackson, Nicholas Miyake, Dan Boneh, and John C. Mitchell. A Browser Plug-In Solution to the Unique Password Problem. http://crypto.stanford.edu/PwdHash/, 2005.
[7] Blake Ross, Collin Jackson, Nicholas Miyake, Dan Boneh, and John C. Mitchell. Stronger Password Authentication Using Browser Extensions. In 14th Usenix Security Symposium, 2005.
[8] YZhang , J.Hong , L.Cranor. "CANTINA :A content based approach to detect phishing websites" in Proceedings of the 16th International Conference on World Wide Web , Banff, AB, Canada pp 639-64S. May OS - 12 2007.
[9] “Investigating the effect of feature selection and dimensionality reduction on phishing website classification problem”, international conference on Next Generation Computing Technologies (NGCT-2015) 2015.
[10] Whois. Available at: http://www.who.is/WHOisDatabase
[11] Christian Ludl, Sean McAllister, Engin Kirda, Christopher Kruegel "On the Effectiveness of Techniques to Detect Phishing Sites" at Proc 4th International Conference, DIMV A 2007 Lucerne, Switzerland, pp 20-39 July 12-13,2007
[12] Keng Siau and Sang Juan Lee "A Review of Data Mining Techniques" in Industrial Management & Data Systems , MCB University Press ,pp 41-46 , 2001.
[13] Kantardzic and Mehmed. "Data Mining.: Concepts, Models, Methods, and Algorithm", John Wiley & Sons, Wiley-IEEE Press, July 2011.
[14] Mustafa AYDIN, Nazife BAYKAL CyDeS, “Feature Extraction and Classification Phishing Websites Based on URL” Cyber Defense and Security Laboratory of METU-COMODO Informatics Institute, Middle East Technical University (METU) Ankara, Turkey.
[15] Christian Ludl, Sean McAllister, Engin Kirda, Christopher Kruegel ,“On the Effectiveness of Techniques to Detect Phishing Sites” Secure Systems Lab, Technical University Vienna.