An Effective Strategy for Identifying Phishing Websites using Class-Based Approach

K. Ruth Ramya, K. Priyanka, K. Anusha, Ch. Jyosthna Devi, Y. A. Siva Prasad

Abstract-This paper presents a novel approach to overcome the difficulty and complexity in detecting and predicting social networking phishing website. We proposed an intelligent resilient and effective model that is based on using A New Class Based Associative Classification Algorithm which is an advanced and efficient approach than all other association and classification Data Mining algorithms. This algorithm is used to characterize and identify all the factors and rules in order to classify the phishing website and the relationship that correlate them with each other. Applying the association rule into classification can improve the accuracy and obtain some valuable rules and information that cannot be captured by other classification approaches. The class label is taken good advantage in the rule mining step so as to cut down the searching space. The proposed algorithm also synchronize the rule generation and classifier building phases, shrinking the rule mining space when building the classifier to help speed up the rule generation.

Key words: Association rule, Classification , Data Mining, Data Set, Phishing ,Pruning, Rule Mining

1.INTRODUCTION

Phishing is a online deception technique in which scam artists uses an e-mail or website to illegitimately obtain confidential information such as user names and passwords. Phishing makes use of spoofed emails that are made to look authentic and purported to be coming from legitimate sources. It is a semantic attack which targets the user rather than the computer. It is a new internet crime . Social networking sites are now a major objective of phishing. The motivation behind this study is to create a resilient and effective method that uses Data Mining algorithm and tools to detect phishing websites. Class Based Associative classification algorithms can be very useful in predicting Phishing websites. It can give us answers about what are the most important social networking phishing websites characteristics and indicators and how they relate with each other. The paper is organized as follows: Section 2 presents the phishing scams, Section 3 presents the literature review, Section 4 presents the case studies of phishing websites, Section 5 presents Phishing Indicators, Section6 presents Class based associative classification, Section 7 Website Phishing Training Data Sets and then conclusions are given in Section 8.

2.Phishing Scams

There are many ways in which someone can use phishing to social engineer someone. For example, someone can manipulate a website address to make it look like you are going to a legitimate website, when in fact you are going to a website hosted by a criminal. The process of phishing [12] involves five steps namely, planning, setup, attack, collection and identity theft and fraud. During the planning stage the phishers decide which business to target and determine how to get e-mail addresses for the customers of that business. They often use the same mass-mailing and address collection techniques as spammers. In the setup stage after they know which business to spoof and who their victims are, the phishers create methods for delivering the message and collecting the data. Most often, this involves e-mail addresses and a web page. The attack stage is the step people are most familiar with - the phisher sends a phony message that appears to be from a reputable source. The collection stage is the one in which phishers record the information entered by victims into Web pages or popup windows. The final stage is the Identity theft and Fraud where the phishers use the information they've gathered to make illegal purchases or otherwise commit fraud as many as a fourth of the victims never fully recover. If the phisher wants to coordinate another attack, he evaluates the successes and failures of the completed scam and begins the cycle again.

Fig 1 represents how the phishing attack is done and by that how the user identity is revealed to the phisher.Phisher by using that confidential information will try to act like the alleged user and thereby breaches the security of the user.

- *K.Anusha is currently pursuing bachelor degree program in computer science and engineering in KL University, India, PH-9490020310. E-mail: anu.korrapati01@gmail.com*

- *K.Priyanka is currently pursuing bachelor degree program in computer science and engineering in KL University, India, PH-9502507607. E-mail: kpriyanka.btech@gmail.com*

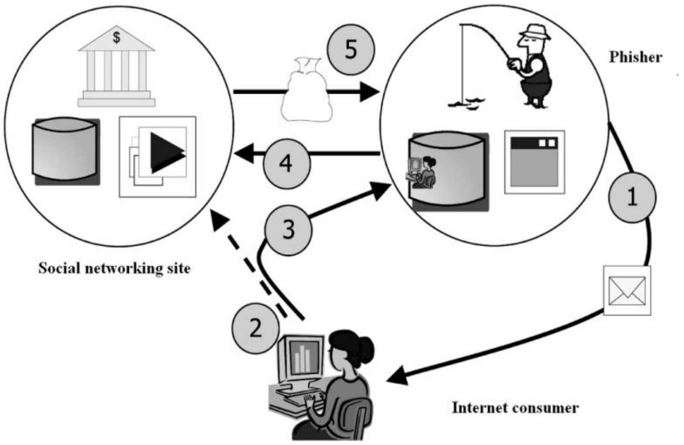

Fig 1: Phishing 1:Phisher sends an email to user 2:User thinks that it comes from original website 3:User enter confidential information 4:Phisher enter the user datails into the original website 5: Phisher extract the user information.

3.Literature Review

A report by Kelly Jackson Higgins estimated that phishing attacks on social networking sites increased more than 240 percent in 2008 compared to the same 2009[7][8], just behind attacks on payment services, which jumped a whopping 285 percent versus the first quarter of '2008'. Moore and Clayton estimated the number of phishing victims by examining web server logs. They estimated that 311,449 people fall for phishing scams annually, costing around 350 million dollars [4] There are several promising defending approaches to this problem reported earlier. One approach to stop phishing is by using Visual or Audio Personalization of E-mail. This approach offers a simple visual or audio mechanism to verify the authenticity of E-mail. The weakness of this approach is that we must physically appear at the institution's office in order to take the picture and must strongly protect the database containing the authentication data(pictures, sound clips, or pass phrases).Second approach to stop phishing is by using Secure Token Authentication. In this type of authentication system,the hardware token provides a one-time password that is valid only for the owner of the token.The token generates a new one-time password with each login, so it does not matter if an attacker obtains the value. The weakness of this approach is that secure tokens do not prevent the user from supplying information that could be used as a proxy for the information to carry out a transaction. In particular, information used to identify a person, such as their mother's maiden name, can still be obtained and could lead to fraudulent activities. Third approach to stop phishing is by using Active Web Monitoring. This approach involves development of the equivalent of "white-list" admissibility tests of trademark and key content. Monitoring service companies deploy agent-based solutions to continuously monitor web content, actively searching for all instances of a client's logo,trademark, or key web content. The weakness of this approach is Time delay between identification and action to eliminate use may still result in numerous thefts of private information and requires active monitoring. Fourth approach to stop phishing is by using Gateway Anti-Spam Filtering. Here anti-spam filtering can block some fraudulent E-mail before it is ever delivered to the user. Phishing E-mails are one particular form of spam. The weakness of this approach is Spam detection is improving, but as spammers constantly change their spamming techniques no solution can be 100% accurate. Due to these imperfections, users may elect to review all suspected spam before deleting it. The user must learn to recognize false positives.Another approach to stop phishing is by using Desktop Privacy Service. In this Commercially available software packages can monitor outgoing web traffic for a user-definable set of data. The data is typically defined to be information that identifies the user, such as names social security numbers, and credit card numbers. If any of that set of data appears in the outgoing packets, the packet is halted until the user confirms that the data should be sent to the true destination, or that the data delivery should be aborted.

4.Case Studies on Phishing Sites

Phishing occurs despite the growing efforts that are taken to educate users and users are still very much susceptible to phishing attacks. Even the display of Extended Validation certification did not decrease the percentage of users who are made victims of phishing attacks.

Case Study 1: Phone Phishing Experiment

For our testing specimen, a group of 50 employees were contacted by female colleagues assigned to lure them into giving away their personal information such as user names and passwords.The results were surprisingly beyond expectations; many of the employees fell for the trick. After conducting friendly conversations with them for some time, our team managed to reduce them into giving away their Internet banking credentials for fake reasons. Some of these lame reasons included checking their privileges and accessibility, or checking the account's integrity and connectivity with the Web server for maintenance purposes, account security and privacy assurance. To assure the authenticity of our request and to give it a social dimensional trend, our team had to contact them repeatedly, perhaps three or four times. Our team managed to deceive 16 out of the 50 employees into giving away their full credentials (user name and password), which represented 32% of the sample. This percentage is considered a high one especially when we know that the victims were staff members of a bank, who are supposed to be highly educated with regard to the risks associated with electronic banking services. A total of eight employees (16%) agreed to give their user name only and refrained from giving away their passwords.

Response to Phone Phishing Experiment	**Number of Employees**
Giving away their full credentials (user name and password)	16
Giving away only their credentials user name without password	8
Refused to reveal their credentials or any kind of information	26
Total	50

Table1 :Phone Phishing Experiment

Case Study 2: Website Phishing

Consider the original website and the phished website of a bank namely, the State Bank of India (SBI) which is involved in e-banking. Unless the user is a known visitor of the site it is not possible for him/her to identify the authentication of the site based on its look and feel. When we take a close look at the two sites some differences can be observed, URL is different - The URL of the original site is **www.onlinesbi.com[9]** and the URL of the phished website is **www.sbionline.com [10]**and Validation of the EV SSL certificate - Extended Validation Secure Sockets Layer (SSL) Certificates are special SSL Certificates that work with high security Web browsers to clearly identify a Web site's organizational identity. Extended Validation (EV) helps you make sure a Web site is genuine and verified. In original websites, the address bar turns green indicating that the site is secured by an EV certificate.

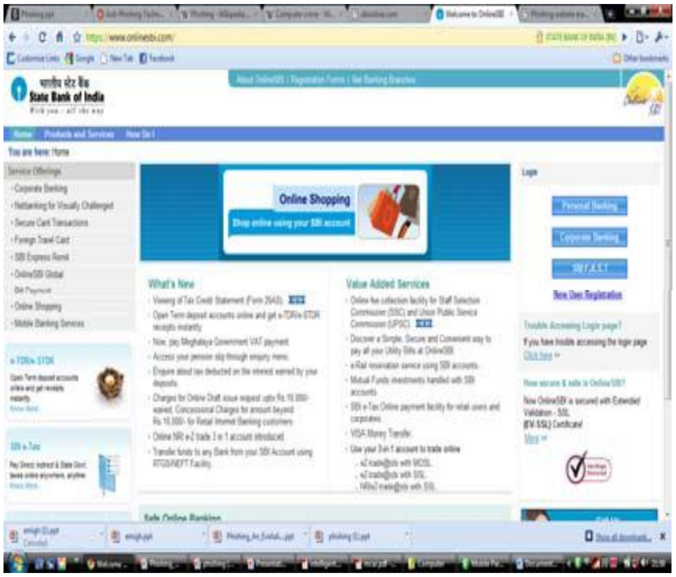

Fig2: Original website of SBI

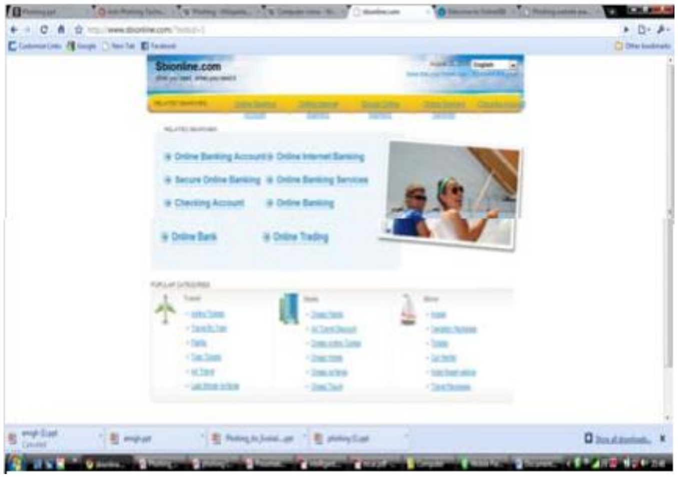

Fig3: Phished website of SBI

Response to Phone Phishing Experiment	**Number of Employees**
Interacted positively(IT Department)	8
Interacted positively(Other Departments)	44
Interacted negatively(Incorrect info)	28
Interacted negatively(no response)	40
Total	120

Table2 :Phishing Website Experiment

5.Phishing Characteristics and Indicators

There are many characteristics and indicators that can distinguish the original legitimate social networking websites from the phishing one. We managed to gather 27 phishing features and indicators and clustered them into six Criteria (URL & Domain Identity, Security & Encryption, Source Code & Java script, Page Style & Contents, Web Address Bar and Social Human Factor), and each criteria has its own phishing components[11].

Criteria	N	Phishing Indicators
URL &Domain Identity	1	Using IP address
	2	Abnormal request URL
	3	Abnormal URL of anchor
	4	Abnormal DNS record
	5	Abnormal URL
Security & Encryption	1	Using SSL certificate(Padlock Icon)
	2	Certificate authority
	3	Abnormal cookie
	4	Distinguished names certificate
Source Code & Javascript	1	Redirect Pages
	2	Straaddling attack
	3	Pharming attack
	4	OnmouseOver to hide the link
	5	Server Form Handler(SFH)
Page Style & Contents	1	Spelling errors
	2	Copy website
	3	Using forms with Submit Button
	4	Using Pop-ups WIndows
	5	Disabling right-click
Web Address Bar	1	Long URL address
	2	Replacing similar char for URL
	3	Adding a prefix or suffix
	4	Using the @ Symbol to confuse
	5	Using hexadecimal char codes
Social Human Factor	1	Emphasis on security
	2	Public generic salutation
	3	Buying time to access accounts

Table3:Phishing Indicators

6.Class Based Associative Classification Algorithm (CACA)

Classification is one of the most important tasks in data mining. Researchers are focusing on designing classification algorithms to build accurate and efficient classifiers for large data sets. Being a new classification method that integrates association rule mining into classification problems,associative classification achieves high classification accuracy, its rules are interpretable and it provides confidence probability when classifying objects which can be used to solve classification problem of uncertainty. Therefore, it becomes a hot theme in recent year. The traditional associative classification algorithms basically have 3 phases: Rule Generation, Building Classifier and Classification as shown in Fig4. Rule Generation employ the association rule mining technique to search for the frequent patterns containing classification rules. Building Classifier phase tries to remove the redundant rules, organize the useful ones in a reasonable order to form the classifier and the unlabeled data will be classified in the third step.

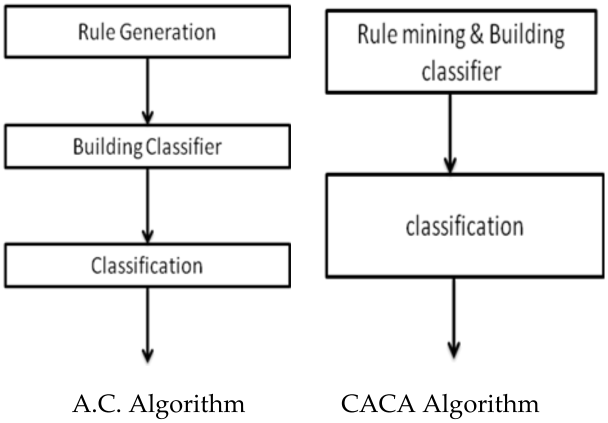

Fig4: Procedures of A.C. Algorithms

However, the drawbacks of associative classification algorithms can be generalized as ,although the associative classification can provide more rules and information, redundant rules may also be included in the classifier which increases the time cost when classifying objects. MCAR[1]determined a redundant rule by checking whether it covers instances in training data set or not . Second, as we know, the rule generation is based on frequent pattern mining in associative classification, when the size of data set grows, the time cost for frequent pattern mining may increase sharply which may be an inherent limitation of associative classification. W. Li, J. Han and J. Pei mine the frequent patterns with FP Growth technique in CMAR which is proved to be very efficient, but extra time should be considered to compute the support and confidence of rules by scanning data set again. In this paper, a class based associative classification algorithm[5] is proposed to solve the difficulty aforementioned. In this algorithm 4 innovations are integrated: 1, use the class based strategic to cut down the searching space of frequent pattern; 2, design a structure call Ordered Rule-Tree to store the rules and their information which may also prepare for the synchronization of the two steps; 3, redefine the compact set so that the compact classifier is unique and not sensitive to the rule reduction; 4, synchronize the rule generation and building classifier phases.

A. Class based rule mining strategy

Given a training data set D with k classes, the principle idea of class based rule mining is to divide the single attribute value set C_{all} for all classes into k smaller ones for every class,that is, to limit the searching in k low dimensional spaces other than a high dimensional one.

B. Ordered Rule Tree Structure (OR-Tree)

To facilitate the synchronization, we design a structure call Ordered Rule Tree under the inspiration of CR-Tree to store and rank rules. It is composed with a tree structure and an ordered list. When a rule r < a_{i1},a_{i2} .. ……a_{il} ,c > satisfying the support and confidence thresholds is generated, attribute values a_{i1},a_{i2},a_{i3},……..a_{iq} will be stored as nodes in this tree according to their frequency in D in descending order. The last node points to an information node storing the rule's information such as class label, support and confidence. Each rule can and only can have one information node. The ordered list is designed to organize all rules in the tree. Each node in the chain points to a certain rule. Nodes pointing to the rules with higher priority are closer to the head node, while those pointing to the rules with lower priority are farther from the head node. When a new non-redundant rule is inserted in the OR-Tree, a new node pointing to this rule will be inserted into a suitable place in the ordered list.

C. Compact Rule Set Redefinition and Pruning Skill

MCAR judge a redundant rule by check whether it cover at least one instance. On one hand, this strategic can not guarantee the removed rule is redundant to the instances n covered by training data set; on the other hand, the reduction should be carry out after all rules are generated and ranked. It is impossible to implement the synchronization with this strategic. However, the definition of compact set and redundant rule in [2], can not ensure the compact classifier is unique and with the same accuracy compared with the original one, which means the classifier and the accuracy changes as the order of rule reduction changes. To overcome these problems, the compact set and redundant rule are redefined in this paper.

Redundant Rule:

Given r_1,r_2,r_3€R,r_2 is redundant if
1.r_1=<$Item_1$,c_k> and r_2=<$Item_1$,c_p>, but r_1>r_2;
2.r_1=<$Item_1$,c_k>, r_2=<$Item_2$,c_p>, $Item_1$<$item_2$; and r_1>r_2
3.r_1=<$Item_1$,c_k>, r_2=<$Item_2$,c_k> $Item_1$<$Item_2$ and r_1>r_2
4.r_1=<$Item_1$,c_k>, r_2=<$Item_2$,c_k> $Item_1$<$Item_2$, r_2>r_1 for r_3=<$Item_3$,c_p>, $Item_1$<$Item_3$, r_3<r_2<r_1

Compact Rule Set:

For rule set R , if R`⊂ R ,any redundant rule r` ∉ R,and R` is unique, then R`is the compact set of R .

Pruning :

For rule r_i=(item,c_i), if supp(r_i)/conf(r_i). (1-conf(r_i))<minsupp , stop mining r_i=($item_k$,c_i,) ,$item_k$ ⊃ item.

D. CACA Algorithm

CACA technically combined the rule generation and the building classifier phases together. Once a new rule is generated, the algorithm visits the OR-Tree partially to recognize its redundancy, stores it in the OR-Tree and ranks it in the rule set. Not only can the synchronization simplify the procedure of associative classification but also apply the pruning skill to shrink the rule mining space and raise the efficiency. The algorithm is design as follow:

(1) CACA first scans the training data set D, stores data in form of vertical representation, counts the frequency of every attribute value a_{ij} and arrange a_{ij} in descending order by frequency. The a_{ij} which is failed to satisfy the minsupp is filtered in this step.

(2) For the remaining attribute values a_{ij} in step (1), Intersect C(a_{ij}) and C(c_n), n = 1,2,…. k . Add a_{ij} into C_n if | D(a_{ij}) ∩D(c_n) | > minsupp .Thus we have k single attribute value sets C_1, C_2,……C_k .

(3)For class c_n, choose a_{i1j1} C_n in accordance with the order, figure out whether rule r=(a_{i1j1}, c_n,) can satisfy minconf (all the elements in single attribute value sets satisfy support threshold) and its redundancy. If it satisfies the threshold and is not a redundant one, it would be inserted and ranked in the OR-Tree. Check whether it satisfies the condition of pruning skill. If yes, let $C_n = C_n \backslash a_{i1j1}$ and repeat (3), else go on with the recursive procedure of mining more detailed rules.

(4) Take an a_{i1j2} € $C_n \backslash a_{i1j1}$, $i_1, \neq i_2$with respect to the frequency order. Judge the satisfaction of minsupp for $r = (a_{i1j1}, a_{i2j2}, c_n)$. Any dissatisfaction leads to a new selection of element, that is,select a_{i3j3} € $C_n \backslash \{a_{i1j1}, a_{i2j2}\}$ and go on with the judgment. Otherwise, if $r = (a_{i1j1}, a_{i2j2}, c_n)$ satisfy theminsupp , check the confidence threshold and redundancy as in step (3). Insert the satisfactory rule in the OR-Tree (or modify the OR-tree when an old rule should be replaced by a new one or an old rule become redundant), rank it and check whether the pruning can be applied here. If the pruning can be carried out here, go back to the upper layer of the recursion. If not, recursively construct Item sets with more attribute values to obtain new rule. When all rules related with c_n and a_{i1j1} is properly handled, the recursion is finished. Then let $C_n = C_n \backslash a_{i1j1}$ repeat (3), until $C_n = \phi$.

(5) Repeat step (3)-(4) until $C_n = \phi$ $n = 1,2,\ldots\ldots\ldots k$.

(6) Classify the unlabeled data by the obtaining classifier.

Model for Detecting Phishing

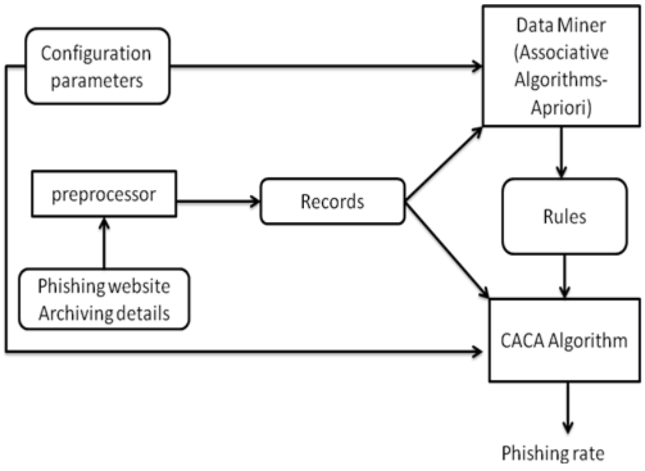
Configuration parameters
Data Miner (Associative Algorithms-Apriori)
preprocessor
Records
Rules
Phishing website Archiving details
CACA Algorithm
Phishing rate

Fig5:Model for Detecting Phishing

Initially all the phishing website details are collected and stored in the phishing website archive. Then it is sent to a preprocessor to convert into machine understandable format. The result is then stored as records in the database. The database also stores configuration parameters (the 27 phishing indicators that are being extracted from the code). Using the data collected in the database, rules are generated to detect the website phishing rate using the CACA algorithm.

7.Website Phishing Training Data Sets

Two publicly available datasets were used to test our implementation: the "phishtank" from the phishtank.com [3] which is considered one of the primary phishing-report collates both the 2007 and 2008 collections. The PhishTank database records the URL for the suspected website that has been reported, the time of that report, and sometimes further detail such as the screenshots of the website, and is publicly available. The Anti Phishing Working Group (APWG) which maintains a "Phishing Archive" describing phishing attacks dating back to September 2007 [6]. A data set of 1006 phishing, suspicious and legitimate social networking websites is used in the study (412 row phishing social networking websites, 288 rows suspicious and 306 row of real social networking websites for the legitimate portion of the data set). In addition, 27 features are used to train and test the classifiers.

8.Conclusion

The associative classification data mining social networking phishing website model showed the significance importance of the phishing website two criteria's (URL & Domain Identity) and (Security & Encryption).According to the characteristic of associative classification, a new class based frequent pattern mining strategic is designed in CACA to cut down the searching space of frequent pattern. OR-Tree structure enables the synchronization of the traditional phases which may not only simplify the associative classification but help to guide the rule generation and speed up the algorithm. And the redefinition of the redundant rule and compact set guarantee the usage of the compact set to help improve the classification efficiency and rule quality won't affect the accuracy of CACA.

References

[1]T., Fadi, c.Peter and Y. Peng, "MCAR: Multi-class Classification based on Association Rule", IEEE International Conference on Computer Systems and Applications ,2005, pp. 127-133.

[2]G. Chen, H. Liu et al, “A new approach to classification based on association rule mining”, Decision Support Systems 42, 2006, pp.674-689.

[3] T. Sharif, “Phishing Filter in IE7,” http://blogs.msdn.com/ie/archive/2005/463204.a spx, , 2006

[4]T. Moore and R. Clayton, "An empirical analysis of the current state of phishing attack and defence", In Proceedings of the Workshop on the Economics of Information Security (WEIS2007)

[5]Zhonghua Tang and Qin Liao"A New Class Based Associative Classification Algorithm"IJAM 2007

[6] Anti-Phishing Working Group. Trends Report , http v/antiphishing.org/apwgrcport_sep_ final.pdf2007 .

[7]http://www.darkreading.com/security/attacks-breaches/218101868/index.html

[8]https://www.markmonitor.com/pr/brandjacking/spring2009/

[9] www.onlinesbi.com retrieved on 26 August,2010

[10] www.sbionline.com retrieved on 26 August,2010

[11] Maher Aburrous, M.A. Hossain, Keshav Dahal,
Fadi Thabtah “Associative Classification Techniques for predicting e-Banking Phishing Websites” IEEE 2010

[12]A.Martin, Na.Ba.Anutthamaa, M.Sathyavathy, Marie Manjari Saint Francois,Dr.Prasanna Venkatesan "A Framework for Predicting Phishing Websites Using
Neural Networks "IJCSI 2011