Predicting Phishing Websites using Classification Mining Techniques with Experimental Case Studies

Maher Aburrous
Dept. of Computing
University of Bradford
Bradford, UK
mrmaburr@bradford.ac.uk

M. A. Hossain
Dept. of Computing
University of Bradford
Bradford, UK
m.a.hossain1@bradford.ac.uk

Keshav Dahal
Dept. of Computing
University of Bradford
Bradford, UK
k.p.dahal@bradford.ac.uk

Fadi Thabtah
MIS Department
Philadelphia University
Amman, Jordan
ffayez@philadelpha.edu.jo

Abstract

Classification Data Mining (DM) Techniques can be a very useful tool in detecting and identifying e-banking phishing websites. In this paper, we present a novel approach to overcome the difficulty and complexity in detecting and predicting e-banking phishing website. We proposed an intelligent resilient and effective model that is based on using association and classification Data Mining algorithms. These algorithms were 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. We implemented six different classification algorithm and techniques to extract the phishing training data sets criteria to classify their legitimacy. We also compared their performances, accuracy, number of rules generated and speed. A Phishing Case study was applied to illustrate the website phishing process. The rules generated from the associative classification model 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 demonstrated the feasibility of using Associative Classification techniques in real applications and its better performance as compared to other traditional classifications algorithms.

Key Words*:* Classification, Association, Data Mining, Fuzzy Logic, Machine Learning.

1. Introduction

Phishing websites is a semantic attack which targets the user rather than the computer. It is a relatively new Internet crime in comparison with other forms, e.g., virus and hacking. The phishing problem is a hard problem because of the fact that it is very easy for an attacker to create an exact replica of a good banking site, which looks very convincing to users. The word phishing from the phrase "website phishing" is a variation on the word "fishing". The idea is that bait is thrown out with the hopes that a user will grab it and bite into it just like the fish. In most cases, bait is either an e-mail or an instant messaging site, which will take the user to hostile phishing websites [7]. The motivation behind this study is to create a resilient and effective method that uses Data Mining algorithms and tools to detect e-banking phishing websites in an Artificial Intelligent technique. Associative and classification algorithms can be very useful in predicting Phishing websites. It can give us answers about what are the most important e-banking phishing website characteristics and indicators and how they relate with each other. Comparing between different Data Mining classification and association methods and techniques is also a goal of this investigation since there are only few studies that compares different data mining techniques in predicting phishing websites. The paper is organized as follows: Section 2 presents the literature review, Section 3 shows the case studies, Section 4 shows data mining phishing approach, Section 5 shows the phishing website methodology of the research, Section 6 shows the utilization of the DM classification techniques, Section 7 reveals the experimental results of implementing the classification data mining techniques in the phishing training data sets and then conclusions and future work are given in Section 8.

2. Literature Review

Despite growing efforts to educate users and create better detection tools, users are still very susceptible to phishing attacks. Unfortunately, due to the nature of the attacks, it is very difficult to estimate the number of people who actually fall victim. A report by Gartner estimated the costs at $1,244 per victim, an increase over

978-0-7695-3984-3/10 $26.00 © 2010 IEEE
DOI 10.1109/ITNG.2010.117

the $257 they cited in a 2004 report [8]. In 2007, 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 [30]. There are several promising defending approaches to this problem reported earlier. One approach is to stop phishing at the email level [2], since most current phishing attacks use broadcast email (spam) to lure victims to a phishing website. Another approach is to use security toolbars. The phishing filter in IE8 [11] is a toolbar approach with more features such as blocking the user's activity with a detected phishing site. A third approach is to visually differentiate the phishing sites from the spoofed legitimate sites. Dynamic Security Skins [3] proposes to use a randomly generated visual hash to customize the browser window or web form elements to indicate the successfully authenticated sites. A fourth approach is two-factor authentication, which ensures that the user not only knows a secret but also presents a security token [4]. However, this approach is a server-side solution. Sensitive information that is not related to a specific site, *e.g.*, credit card information and SSN (Social Security Number), cannot be protected by this approach either [12]. Many industrial antiphishing products use toolbars in Web browsers, but some researchers have shown that security tool bars don't effectively prevent phishing attacks. Authors in [3] proposed a scheme that utilises a cryptographic identity-verification method that lets remote Web servers prove their identities. However, this proposal requires changes to the entire Web infrastructure (both servers and clients), so it can succeed only if the entire industry supports it. In [9], the authors proposed a tool to model and describe phishing by visualizing and quantifying a given site's threat, but this method still wouldn't provide an antiphishing solution. Another approach is to employ certification, e.g., Microsoft spam privacy [28]. A recent and particularly promising solution was proposed in [6], which combines the technique of standard certificates with a visual indication of correct certification. A variant of web credential is to use a database or list published by a trusted party, where known phishing web sites are blacklisted. For example Netcraft, Websense and Cloudmark antiphishing toolbars [29], [10], prevents phishing attacks by utilising a centralized blacklist of current phishing URLs. The weaknesses of this approach are its poor scalability and its timeliness. The typical technologies of antiphishing from the user interface aspect are done by [3] and [12]. They proposed methods that need Web page creators to follow certain rules to create Web pages, either by adding dynamic skin to Web pages or adding sensitive information location attributes to HTML code. However, it is difficult to convince all Web page creators to follow the rules [5]. In [5], [9], the visual similarity of Web pages is oriented, and the concept of visual approach to phishing detection was first introduced. The Passpet system, created by Yee et al. in 2006, uses indicators so that users know they are at a previously-trusted website [13]. Since all of these proposals require the use of complicated third-party tools, its unclear how many users will actually benefit from them. The newest version of Microsoft's Internet Explorer supports Extended Validation (EV) certificates, coloring the URL bar green and displaying the name of the company. However, a recent study found that EV certificates did not make users less fall for phishing attacks [25].

3. Case Studies

3.1. Case Study: Phone Phishing Experiment

For our testing specimen, and after taking all the necessary authorization and approval from the management, a group of 50 employees were contacted by female colleges assigned to lure them into giving away their personal ebanking accounts user name and password (through social and friendly phone conversation with a deceiving purpose in mind). The results were beyond expectations; many of the employees fell for the trick. After conducting friendly conversation with them for some time, our team managed to seduce them into giving away their internet banking credentials for fake reasons. Some of these lame reasons included checking their privileges and accessibility, or for checking its integrity and connectivity with the web server for maintenance purposes, account security and privacy assurance…etc. To assure the authenticity of our request and to give it a social dimensional trend, our team had to contact them repeatedly for about three or four time. As shown in table 1, our team managed to deceive 16 out of the 50 employees to give away their full e-banking credentials which represented 32% of the sample. This percentage is considered a high one especially when we know that the victims were staff members of Jordan Ahli Bank, who are supposed to be highly educated with regard to the risks of electronic banking services.

Table 1. Phone Phishing Experiment

Response to Phone Phishing	No. of Emp.
Giving away their full ebanking credentials(user name & Password)	16
Giving away only their ebanking user name without password	8
Refused to reveal their credentials	26
Total	**50**

A total of 16% (8 employees) agreed to give their user name only and refrained from giving away their passwords under any circumstances or excuses what so ever. The remaining 52% (26 employees) were very cautious and declined to reveal any information regarding their credentials over the phone. An overview of the results reveals the high risk of social engineering security factor. Social engineering constitutes a direct internal threat to e-banking web services since its hacks directly into the accounts of e-bank customers. The results also show the direct need to increase the awareness of customers not to fall victims of this kind of threat that can lead to devastating results.

3.2. Case Study: Website Phishing Experiment

We engineered a website for phishing practice and study. The website was an exact replica of the original Jordan Ahli Bank website www.ahlionline.com.jo designed to trap users and induce them by targeted phishing email to submit their credentials (username and password). The specimen was inclusive of our colleagues at Jordan Ahli Bank after attaining the necessary authorizations from our management. We deliberately put lots of known phishing features and factors when creating the faked website in order to measure the user's awareness of these kinds of risk. For example, using IP address instead of domain name, http instead of https, poor design, spelling errors, absence of SSL padlock icon and phony security certificate. We targeted 120 employees with our deceiving phishing email, informing them that their e-banking accounts are at the risk of being hacked and requested them to log into their account through fake link attached to our email using their usual customer ID and password to verify their balance and then log out normally. As shown in table 2, The website successfully attracted 52 out of the 120-targeted employees representing 44%, who interacted positively by following the deceiving instructions and submitting their actual credentials (customer ID, Password).

Table 2. Phishing Website Experiment

Response to Phishing Website	No. of Emp.
Interacted positively (IT Department)	8
Interacted positively (Other Departments)	44
Interacted negatively (Incorrect info)	28
Interacted negatively (No response)	40
Total	**120**

Surprisingly IT department employees and IT auditors constituted 8 out of the 120 victims representing 7%, which shocked me, since we expected them to be more alert than others. From other departments 44 employees of the 120-targeted employee’s victims representing 37%, fell into the trap and submitted their credentials without any hesitation. The remaining 68 out of 120 representing 56% were divided as follows: 28 employees supplied incorrect info, which seems to indicate a wary curiosity representing 23%; and 40 employees, received the email, but did not respond at all representing 33%. The results clearly indicate that target phishing factor is extremely dangerous since almost half of the employees who responded were victimized; particularly, trained employees such as those of IT Department and IT Auditors. Increasing the awareness of all users of e-banking regarding this risk factor is highly recommended.

4. Phishing Data Mining Approach

4.1. Phishing Characteristics and Indicators

From our previous phishing case studies and experiments 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. For example, URL & Domain Identity Criteria has five phishing indicator components (Using IP address, abnormal request URL, abnormal URL of anchor, abnormal DNS record and abnormal URL).
The full list is shown in table 3 which is used later on our analysis and methodology study.

Table 3. Main Phishing Indicators With Its Criteria

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 & Java script	1	Redirect pages
	2	Straddling attack
	3	Pharming attack
	4	OnMouseOver to hide the Link
	5	Server Form Handler (SFH)
Page Style & Contents	1	Spelling errors
	2	Copying 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

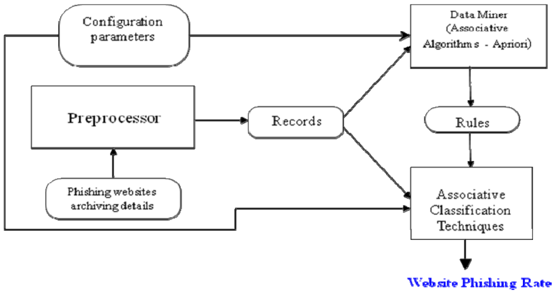
4.2. Why use Data Mining?

Data Mining has been described as "the nontrivial extraction of implicit, unknown, and potentially useful information from large data sets [18]. It is a powerful new technology to help researchers focus on the important information in their data archive. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions [19].

5. Phishing Website Methodology

5.1. Data Mining Techniques

We utilized data mining classification and association rule approaches in our new e-banking phishing website detection model as shown in figure 1 to find the most important phishing features and significant patterns of phishing characteristic or factors in the e-banking phishing website archive data. Particularly, we used a number of different existing data mining association and classification techniques including JRip [21], PART [21], PRISM [22] and C4.5 [23], CBA [20], MCAR [26] algorithms to learn and to compare the relationships of the different phishing classification features and rules. The experiments of C4.5, RIPPER, PART and PRISM algorithms were conducted using the *WEKA* software system [16], which is an open java source code for the data mining community that includes implementations of different methods for several different data mining tasks such as classification, association rule and regression. CBA and MCAR experiments were conducted using an implementation version provided by the authors of [20], [26]. We have chosen these algorithms based on the different strategies they use to generate the rules and since their learnt classifiers are easily understood by human.



Configuration parameters
Data Miner (Associative Algorithms - Apriori)
Preprocessor
Records
Rules
Phishing websites archiving details
Associative Classification Techniques
Website Phishing Rate

Figure 1. AC Model for Detecting Phishing Websites

We used two web access archives, one from APWG archive [1] and one from Phishtank archive [24]. We managed to extract the whole 27 phishing security features and indicators and clustered them to its 6 corresponding criteria as mentioned before in table 3.

5.2. Website Phishing Training Data Sets

Two publicly available datasets were used to test our implementation: the "phishtank" from the phishtank.com [24] 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 [2]. A data set of 1006 phishing, suspicious and legitimate e-banking websites is used in the study (412 row phishing e-banking websites, 288 rows suspicious and 306 row of real e-banking websites for the legitimate portion of the data set). In addition, 27 features are used to train and test the classifiers. We used a series of short scripts to programmatically extract the above features, and store these in an excel sheet for quick reference. Our goal is to gather information about classifying and categorizing of all different e-banking phishing attacks techniques. By thoroughly investigating these phishing attacks we've created a data set containing information regarding what different techniques have been used and how it can be predicted.

5.3. Mining e-Banking Phishing Challenges

The age of the dataset is the most significant problem, which is particularly relevant with the phishing corpus. E-banking Phishing websites are short-lived, often lasting only in the order of 48 hours. Some of our features can therefore not be extracted from older websites, making our tests difficult. The average phishing site stays live for approximately 2.25 days [14]. Furthermore, the process of transforming the original e-banking phishing website archives into record feature row data sets is not without error. It requires the use of heuristics at several steps. Thus high accuracy from the data mining algorithms cannot be expected. However, the evidence supporting the golden nuggets comes from a number different algorithms and feature sets and we believe it is compelling [15].

6. DM Classification Techniques

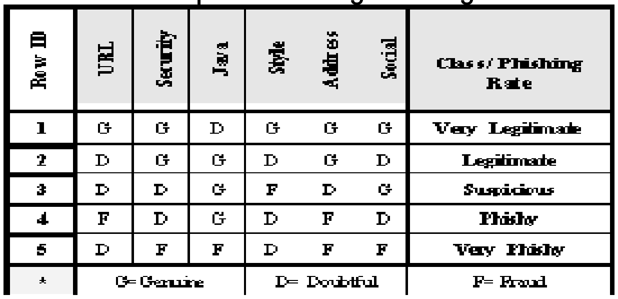
6.1. Classification Algorithm

We utilizes six different common DM classification algorithms (C4.5, JRip, PART, PRISM, CBA and MCAR). Our choice of these methods is based on the different strategies they used in learning rules from data sets [17]. The C4.5 algorithm [23] employs divide and conquer approach, and the RIPPER algorithm uses separate and conquer approach. The choice of PART algorithm is based on the fact that it combines both approaches to generate a set of rules. It adapts separate-and-conquer to generate a set of rules and uses divide-and-conquer to build partial decision trees. PRISM is a classification rule which can only deal with nominal attributes and doesn't do any pruning. It implements a top-down (general to specific) sequential-covering algorithm that employs a simple accuracy-based metric to pick an appropriate rule antecedent during rule construction. CBA algorithm employs association rule mining [20] to learn the classifier and then adds a pruning and prediction steps. Finally, MCAR algorithm consists of two phases: rules generation and a classifier builder. In the first phase, MCAR scans the training data set to discover frequent single items, and then recursively combines the items generated to produce items involving more attributes. MCAR then generates, ranks and stores the rules. In the second phase, the rules are used to generate a classifier by considering their effectiveness on the training data set. This results in a classification approach named associative classification [26] [27]. MCAR utilizes database coverage pruning to decrease the number of rules. Since without adding constraints on the rule discovery, the very large numbers of rules, make humans unable to understand classifier. This pruning technique tests the generated rules against the training data set, and only high quality rules that cover at least one training instance not considered by other higher ranked rules are kept for later classification.

6.2. AC Phishing Approach (MCAR Model)

Associative Classification is a special case of association rule mining in which only the class attribute is considered in the rule's right-hand-side (consequent), for example A, B→Y, Then A, B must be input items attributes and Y must be the output class attribute. The attribute values for all our input items which represent the six e-banking phishing features and criteria (URL & Domain Identity, Security & Encryption, Source Code & Java script, Page Style & Contents, Web Address Bar, and Social Human Factor) ranged between three fuzzy set values (Genuine, Doubtful and Legitimate) which we measured before in our previous paper using Fuzzy Logic [31] taking into consideration all the input fuzzy variables for all criteria different components as shown in Table 3. The output class attribute of our phishing website rate is one of these values (*Very Legitimate, Legitimate, Suspicious, Phishy or Very Phishy*). Example of the training data sets to be classified is shown in Table 4.

Table 4. Example of Training Phishing Data Sets



Row ID	URL	Security	Java	Style	Address	Social	Class / Phishing Rate
1	G	G	D	G	G	G	Very Legitimate
2	D	G	G	D	G	D	Legitimate
3	D	D	G	F	D	G	Suspicious
4	F	D	G	D	F	D	Phishy
5	D	F	F	D	F	F	Very Phishy
*	G= Genuine			D= Doubtful			F= Fraud

To derive a set of class association rules from the training data set, it must satisfy certain user-constraints, i.e support and confidence thresholds. In association rule mining, any item that passes *MinSupp* is known as a frequent item. A sample of the 22 best classification rules generated from the MCAR (*Multi Class Classification based on Association Rule*) algorithm which had an accuracy of 88.4 % and error rate of 12.622 % is shown below.

Rule 1: Social_Human_Factor = Fraud
Web_Address_Bar = Fraud
Page_Style_&_Contents = Doubtful
-> class = Phishing
Rule 16: Web_Address_Bar = Genuine
Security_&_Encryption = Doubtful
URL_Domain_Identity = Doubtful
-> class = Legitimate
Rule 22: Social_Human_Factor = Genuine
Page_Style_&_Contents = Doubtful
-> class = Suspicious

We recorded the prediction accuracy and the number of rules generated by the traditional classification algorithms and the new associative classification approaches we used in Table 5 and 6 respectively. Experiments were conducted using stratified ten-fold cross-validation. Error rate comparative chart is shown in figure 2.

Table 5. Results From Weka four Classifiers

	C4.5	P.A.R.T.	JRip	PRISM
Test Mode	10 FOLD CROSS VALIDATION			
Attributes	URL DOMAIN IDENTITY SOURCE CODE & JAVA WEB ADDRESS BAR		SECURITY & ENCRYPTION PAGE STYLE & CONTENTS SOCIAL HUMAN FACTOR	
No.of Rules	57	38	14	155
Correctly Classified	848 (84.2 %)	869 (86.3 %)	818 (81.3%)	855 (84.9%)
Incorrectly Classified	158 (15.7%)	137 (13.6%)	188 (18.6 %)	141 (14.0%)
Instances	1006	1006	1006	1006

Table 6. Results from CBA and MCAR Classifiers

	CBA	MCAR
Num of Test Case	1006	1006
Correct Prediction	873	886
Error Rate	13.452%	12.622%
Min Sup	20.000%	20.000%
Min Conf	100.000%	100.000%
Rule Limit	80000	80000
Level Limit	6	6
Number of rules	15	22

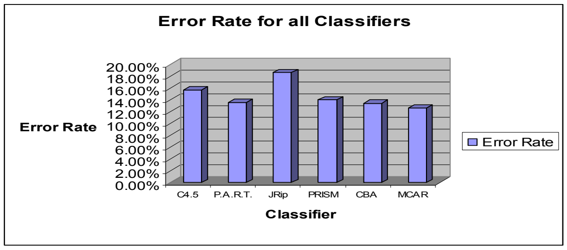


Figure 2. Error Rate For All Classifiers.

7. Experimental Results

Associative classifiers produce more accurate classification models and rules than traditional classification algorithms. The experiments demonstrate the feasibility of using Associative Classification techniques in real applications involving large databases. It showed that there is a significant relation between the two phishing website criteria's (*URL & Domain Identity*) and (*Security & Encryption*) for identifying e-banking phishing website, taking into consideration its characteristic association and relationship with each others (Ex.: Conflict relationship of using SSL certificate VS the abnormal URL request). Also we found insignificant trivial influence of the (*Page Style & Content*) criteria along with (*Social Human Factor*) criteria for identifying e-banking phishing websites. The number of rules extracted from the PRISM algorithms is the highest since it doesn't do any pruning process. Finally, the experiment results showed that the proposed associative classification MCAR algorithm outperformed all other traditional classification in terms of accuracy and speed (Err. Rate: 12.622 %) since it requires only one phase to discover frequent items and rules.

8. Conclusion and Future Work

E-banking phishing website model based on classification data mining showed the significance importance of the phishing website two criteria's (URL & Domain Identity) and (Security & Encryption) in the final phishing detection rate, and also showed the insignificant trivial influence of some other criteria like 'Page Style & content' and 'Social Human Factor' in the final phishing rate. The rules generated from the associative classification model showed the correlation and relationship between some of their characteristics which can help us in building phishing website detection system. The experiments demonstrate the feasibility of using Associative Classification techniques in real applications involving large databases and its better performance as compared to other traditional classification algorithms, [Ex: MCAR (Err. Rate : 12.622 %)]. As for future work, we want to use different pruning methods like lazy pruning [26] which discards rules that incorrectly classify training instances and keeps all other rules to be used by MCAR associative classification technique in order to minimize the size of the resulting classifiers and to experimentally measure and compare the effect of these different pruning on the final result.

9. References

[1] Anti-Phishing Working Group. Phishing Activity Trends Report,http://antiphishing.org/apwg_report_final.pdf. 2007.

[2] B. Adida, S. Hohenberger and R. Rivest, "Lightweight Encryption for Email," USENIX Steps to Reducing Unwanted Traffic on the Internet Workshop (SRUTI), 2005.

[3] R. Dhamija and J.D. Tygar, "The Battle against Phishing: Dynamic Security Skins," Proc. Symp. Usable Privacy and Security, 2005.

[4] FDIC., "Putting an End to Account-Hijacking Theft," http://wwfdic.gov/consumers/id/identity_theft.pdf, 2004.

[5] A. Y. Fu, L. Wenyin and X. Deng, " Detecting Phishing Web Pages with Visual Similarity Assessment Based on Earth Mover's Distance (EMD) ," IEEE transactions on dependable and secure computing, vol. 3, no. 4, 2006.

[6] A. Herzberg and A. Gbara, "Protecting Naive Web Users," Draft of July 18, 2004.

[7] L. James**,** "Phishing Exposed," Tech Target Article by: Sunbelt software, searchexchange.com, 2006.

[8] GARTNER, INC. Gartner Says Number of Phishing E-Mails Sent to U.S. Adults Nearly Doubles in Just Two Years. http://www.gartner.com/it/page.jsp?id=498245, November 9 2006.

[9] W. Liu, X. Deng, G. Huang and A. Y. Fu, "An Antiphishing Strategy Based on Visual Similarity Assessment," Published by the IEEE Computer Society 1089-7801/06 , INTERNET COMPUTING IEEE, 2006.

[10] Y. Pan and X. Ding, "Anomaly BasedWeb Phishing Page Detection," Proceedings of the 22nd Annual Computer Security Applications Conference (ACSAC'06), Computer Society, 2006.

[11] T. Sharif, "Phishing Filter in IE7," http://blogs.msdn.com/ie/archive/2005/463204.aspx, , 2006.

[12] M. Wu, R. C. Miller and G. Little, "Web Wallet: Preventing Phishing Attacks by Revealing User Intentions," MIT Computer Science, 2006.

[13] K. P. Yee and K. Sitaker. "Convenient password management and phishing protection". In SOUPS '06: Proceedings of the Second Symposium on Usable Privacy Security (New York, USA, 2006), ACM Press, pp. 32–43.

[14] FDIC, Tech. Rep., "Putting an end to account-hijacking identity theft", Dec.2004.[Online]. Available: http://www.fdic.gov/consumers/idtheftstudy/identity theft.pdf.

[15] Ian Fette, Norman Sadeh and Anthony Tomasic, "Learning to Detect Phishing Emails", Institute for Software Research International, CMU-ISRI-06-112, June 2006.

[16] WEKA - University of Waikato, New Zealand, EN, 2006: "Weka - Data Mining with Open Source Machine Learning Software in Java";http://www.cs.waikato.ac.nz/ ml/weka (2006/01/31).

[17] Sebastian Misch, "Content Negotiatio in Internet Mail", Diploma Thesis, University of Applied Sciences Cologne, Mat.No.: 7042524, February 2006.

[18] Kantardzic and Mehmed. "*Data Mining: Concepts, Models, Methods, and Algorithms*.", John Wiley & Sons. ISBN 0471228524. OCLC 50055336, 2003.

[19] U.M. Fayyad, "Mining Databases: Towards Algorithms for Discovery," Data Eng. Bull., vol. 21, no. 1, pp. 39-48, 1998.

[20] Bing Liu, Wynne Hsu, Yiming Ma, "Integrating Classification and Association Rule Mining." *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98, Plenary Presentation)*, New York, USA, 1998.

[21] I.H. Witten and E. Frank, "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, CA, 2005.

[22] J. Cendrowska., "*PRISM: An algorithm for inducing modular rule",* International Journal of Man-Machine Studies (1987), Vol.27, No.4, pp.349-370.

[23] J. R. Quinlan, "Improved use of continuous attributes in c4.5", Journal of Artificial Intelligence Research, 4:77-90, 1996.

[24] http://www.phishtank.com/phish_archive.php

[25] C. Jackson, D. Simon, D. Tan, and A. Barth, "An evaluation of extended validation and picture-in-picture phishing attacks". In Proceedings of the 2007 Usable Security (USEC'07). http://www.usablesecurity.org/papers/jackson.pdf.

[26] 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.

[27] F.Thabtah, P. Cowling and Y. Peng, "*A new multi-class,multi-label associative classification approach*".The 4th International Conference on Data Mining(ICDM'04), Brighton, UK, 2004.

[28] Microsoft, "www.microsoft.com/twc/privacy/spam", 2004.

[29] Netcraft, "http://toolbar.netcraft.com/", Dec 2004.

[30] 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)

[31] M. R. Aburrous, M. A. Hossain and K. P. Dahal, "Intelligent Phishing Website Detection System using Fuzzy Techniques", *IEEE ICTTA 2008, 7-11 April* , Damascus, Syria.

[32] Tom N. Jagatic, Nathaniel A. Johnson, Markus Jakobsson and F. Menczer. "Social Phishing Commun". ACM,50(10): 94-100, 2007.