Implementing a Web Browser with Phishing Detection Techniques

Er. Monika Bansal
M.tech CSE , Research Scholar
GKU, Talwandi Sabo (Bathinda)
Bansal.monics@gmail.com

Dr. Dinesh Kumar
Associate Professor, Department of CSE
GKU, Talwandi Sabo (Bathinda)
kdinesh.gku@gmail.com

Abstract

Phishing is the blend of social building and specialized adventures intended to persuade a casualty to give individual data, as a rule for the fiscal addition of the aggressor. Phishing has turn into the most well known practice among the culprits of the Web. Phishing assaults are turning out to be more successive and modern. The effect of phishing is extraordinary and noteworthy since it can include the danger of wholesale fraud and money related misfortunes. Phishing tricks have turn into an issue for internet managing an account and e-trade clients. In this paper we propose a novel way to deal with identify phishing assaults. We actualized a model web program which can be utilized as an operators and procedures every arriving email for phishing assaults. Utilizing email information gathered more than a period time we show information that our methodology has the capacity distinguish more phishing assaults than existing schemes.Our methodology gives comparative exactness to boycotting methodologies (96%), with the point of preference that it can order zero-day phishing assaults and focused on assaults against littler locales, (for example, corporate intranets). A key commitment of this paper is that it incorporates an execution examination and a system for making utilization of PC vision systems in a handy manner.

9 772278 684800 03
© JRPS International Journal for Research Publication & Seminar

I. INTRODUCTION

Phishing is an attack that makes Internet users reveal their personal information to unauthorised party. Most phishing attacks start when users receive fake emails asking them to click a URL (link) to update their accounts' information. Once clicked, this URL will deliver the user to a fake website where he/she will most probably lose control over its account information. According to Anti-Phishing Working Group report, the number of URLs which were used to host phishing attacks has increased from 164,023 in the first quarter of 2012 to 175,229 in the second quarter of the same year [1]. To detect phishing emails, it is important to choose the right detection feature(s). Among the available various antiphishing solutions, there is a considerable number of features which have been suggested to best classify ham (legitimate) and phishing emails. However, in many cases, these features are inappropriately chosen. This is because they are selected based on the author's intuition about their effectiveness in email classification process [2]. This work presents a method to choose the most efficient feature in detecting phishing emails. The importance of the selected feature is determined by calculating its Effectiveness Metric (EM) value based on three criteria which derived based on, and related to three general aspects of email. These three aspects of email are, email's sender, email's content, and email's receiver.

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.

II. FEATURE SELECTION PROCESS

The process of calculating the EM values of the Keywords and URLs features. EM values of these two features were calculated in order to compare their efficiency in detecting phishing emails. Since the email's body is the foremost part that users are concerning about and paying attention to, the features extracted from this part of the email are assumed to have higher importance in detecting phishing attempts than the features extracted from email's header part, and many of cues that influence user's decision about email(s) in question can be found in the email's body part [7]. The Body_no_FunctionWords feature (used in [2], and which is called the Keywords feature in this study) is a content-based feature which has not listed in Table II above. However, this feature has shown its importance in the experiment conducted in [2], it was ranked as the 1st, 16th, and 13th best amongst 40 features in three combinations of the three analyzed datasets in that experiment. In this work, we have focused on the Keywords and the URLs features which are extracted from the email's body part because these two features have a considerable importance . The Keywords feature was used to count occurrences of the selected 18 keywords in the two types of analyzed emails, whereas the URLs feature was used to count the presence and absence occurrences of fake URLs' indications in these emails. A. Feature's Effectiveness Criteria By considering email's sender, email's content, and email's receiver aspects, we have derived three effectiveness criteria which used in calculating the EM values of the Keywords and URLs features and hence to compare their efficiency in detecting phishing emails. Each of these three criteria has given a 1 3⁄ of the effectiveness weight (effectiveness/3). Table III shows these effectiveness criteria and to which aspect of the email each criterion is relate to.

III. 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. 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 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 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.

IV. SCOPE

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. Our Objectives are as follows:

- Data mining tool is used to analyze the email phishing detection from websites.
- The implemented work to extract the phishing training data sets criteria to classify their legitimacy with six different classification algorithm and techniques.
- we also compared their performances, accuracy, number of rules generated and speed.
- The proposed work is to selecting more efficient feature in detecting phishing emails.
- The Effectiveness Metric (EM) values of email classification features are implemented.

V. METHODOLOGY

This is to detect the phishing from the website. It is based upon weka tool. There is different information about the dataset. From the previous study of phishing detection we managed to gather 27 phishing features and indicators and clustered them into six Criteria (URL &

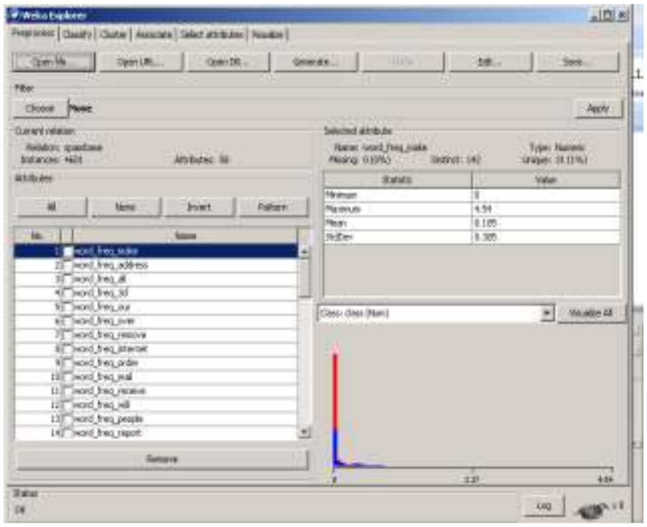
Domain Identity, Security & Encryption, Source Code & Java script, Page Style & Contents, Web Address Bar and Social Human Factor), and each criteria has its own phishing components. For example, URL & Domain Identity Criteria has five phishing indicator components (Using IP address, abnormal request URL, abnormal URL of anchor, abnormal DNS record and abnormal URL). We used a number of different existing data mining association and classification techniques including JRip, PART, ZeroR algorithms to learn and to compare the relationships of the different phishing classification features and rules. All experiments are conducted using the WEKA software system , 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.

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

The proposed Steps for research work:

Step 1: Start the weka tool.
Step 2: Browse the dataset for preprocessing.
Step 3: Select the attribute with different attribute selection.
Step 4: Keep the selected attribute and remove the unselected attribute.
Step5: Classify the selected attribute with different classifier.
Step 6: Analyze the different values after the classification.
Step 7: Visualize the resulted graph with different values.
Step 8: Repeat the step 3 to step 7 for different classifiers.
Step 9: Stop

RESULTS



Relation: spambase
Instances: 4601   Attributes: 58

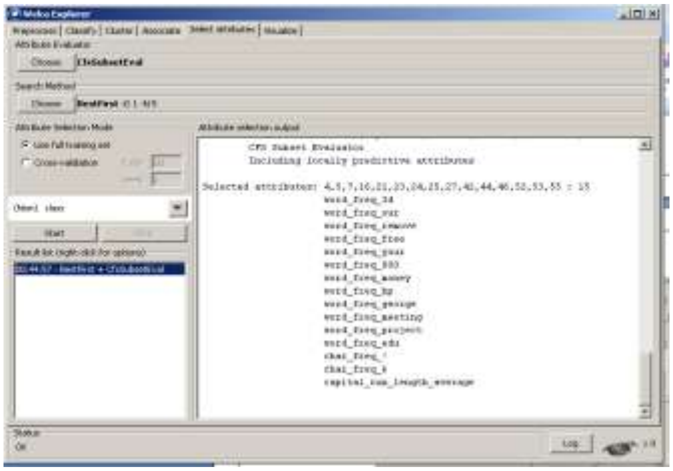
Name: word_freq_make   Type: Numeric
Missing: 0 (0%)   Distinct: 142   Unique: 38 (1%)

Statistic	Value
Minimum	0
Maximum	4.54
Mean	0.105
StdDev	0.305

No.	Name
1	word_freq_make
2	word_freq_address
3	word_freq_all
4	word_freq_3d
5	word_freq_our
6	word_freq_over
7	word_freq_remove
8	word_freq_internet
9	word_freq_order
10	word_freq_mail
11	word_freq_receive
12	word_freq_will
13	word_freq_people
14	word_freq_report

Class: class (Nom)

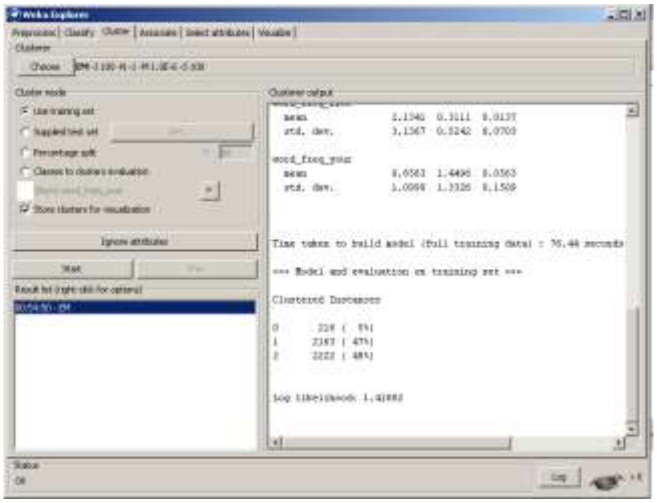
Fig. 1 Displaying data set attributes



Attribute Evaluator: CfsSubsetEval
Search Method: BestFirst -D 1 -N 5
(Nom) class
Result list: 00:44:57 - BestFirst + CfsSubsetEval

```
        CFS Subset Evaluation
        Including locally predictive attributes

Selected attributes: 4,5,7,16,21,23,24,25,27,42,44,46,52,53,55 : 15
                     word_freq_3d
                     word_freq_our
                     word_freq_remove
                     word_freq_free
                     word_freq_your
                     word_freq_000
                     word_freq_money
                     word_freq_hp
                     word_freq_george
                     word_freq_meeting
                     word_freq_project
                     word_freq_edu
                     char_freq_!
                     char_freq_$
                     capital_run_length_average
```

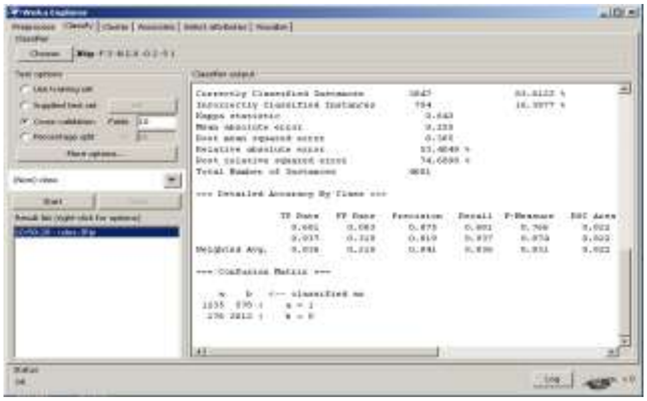
Fig 2 Apply the zerorip classifier



Clusterer: EM -I 100 -N -1 -M 1.0E-6 -S 100
☑ Store clusters for visualization
Result list: 00:54:50 - EM

```
  mean                    2.1346  0.3111  0.0137
  std. dev.               3.1367  0.5242  0.0703

word_freq_your
  mean                    0.6563  1.4496  0.0363
  std. dev.               1.0068  1.3326  0.1509



Time taken to build model (full training data) : 76.44 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       228 (  5%)
1      2183 ( 47%)
2      2222 ( 48%)


Log likelihood: 1.41882
```

Fig 3 Apply effective matrix classifier


Weka Explorer
Preprocess | Classify | Cluster | Associate | Select attributes | Visualize
Choose
Start
Stop
Log

Classifier: JRip -F 3 -N 2.0 -O 2 -S 1

Test options: ○ Use training set ○ Supplied test set ● Cross-validation Folds 10 ○ Percentage split % 66

(Nom) class

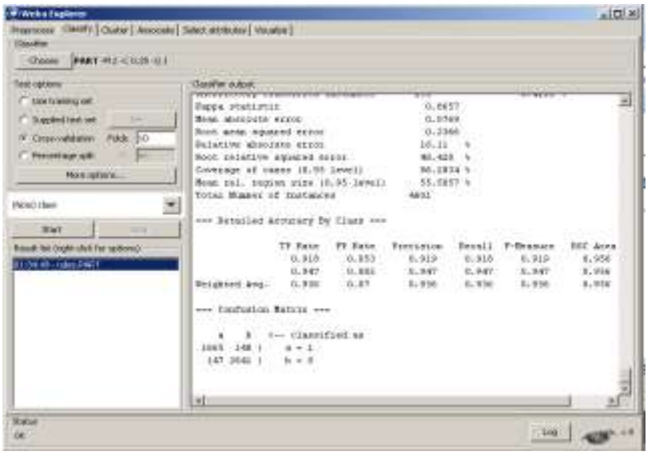
Result list (right-click for options): rules.JRip

Classifier output

```
Correctly Classified Instances        3847               83.6123 %
Incorrectly Classified Instances       754               16.3877 %
Kappa statistic                          0.643
Mean absolute error                      0.233
Root mean squared error                  0.366
Relative absolute error                 53.4849 %
Root relative squared error             74.6888 %
Total Number of Instances             4601

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area
                 0.681     0.063     0.875      0.681     0.766      0.822
                 0.937     0.319     0.819      0.937     0.874      0.822
Weighted Avg.    0.836     0.218     0.841      0.836     0.831      0.822

=== Confusion Matrix ===

    a    b   <-- classified as
 1235  578 |    a = 1
  176 2612 |    b = 0
```

Status
OK

Fig 4 Apply Jrip Classifier


Weka Explorer
Preprocess | Classify | Cluster | Associate | Select attributes | Visualize
Choose
Start
Stop
Log

Classifier: PART -M 2 -C 0.25 -Q 1

Test options: ○ Use training set ○ Supplied test set ● Cross-validation Folds 10 ○ Percentage split % 66

(Nom) class

Result list (right-click for options): rules.PART

Classifier output

```
Kappa statistic                          0.8657
Mean absolute error                      0.0769
Root mean squared error                  0.2366
Relative absolute error                 16.11    %
Root relative squared error             48.428   %
Coverage of cases (0.95 level)          96.2834 %
Mean rel. region size (0.95 level)      55.5857 %
Total Number of Instances             4601

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area
                 0.918     0.053     0.919      0.918     0.919      0.956
                 0.947     0.082     0.947      0.947     0.947      0.956
Weighted Avg.    0.936     0.07      0.936      0.936     0.936      0.956

=== Confusion Matrix ===

    a    b   <-- classified as
 1665  148 |    a = 1
  147 2641 |    b = 0
```

Status
OK

Fig 5 Apply Part Classifier

Table Using Jrip Classifier

Weighted Average	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.681	0.063	0.875	0.681	0.766	0.822	1
	0.937	0.319	0.819	0.937	0.874	0.822	0
	0.836	0.218	0.841	0.836	0.831	0.822	

Table Using Part Classifier

Weighted Average	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.681	0.068	0.867	0.681	0.763	0.858	1
	0.932	0.319	0.818	0.932	0.871	0.858	0

Table Using ZeroR Classifier

Weighted Average	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0	0	0	0	0	0.499	1
	1	1	0.606	1	0.755	0.499	0
	0.606	0.606	0.367	0.606	0.457	0.499	

CONCLUSION

In this paper, we have presented an approach to detect phishing emails using link based features. The contribution of the work mainly consists of the usage of features visible links, invisible links and unmatched urls. The proposed algorithm used in conjunction with the proposed prototype of web browser will help the user to get notified about the possible phishing attacks and prevent them from opening the suspicious websites. The word phishing from the phrase "website phishing" is a variation on the word "fishing". The idea behind that is 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 phishing websites. The most significant problem, which is particularly relevant with the phishing corpus. 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. Phishing websites is a semantic attack which targets the user rather than the computer. It is a relatively new Internet crime compare to other forms, i.e., virus and hacking. The Data mining tool is used to implement the email phishing detection from websites. The implemented work is to extract the phishing training data sets criteria to classify their legitimacy with different classification algorithm and techniques.

.

References:

[1] Zhang, J., et. al,. A. Modified logistic regression: An approximation to SVM and its applications in large-scale text categorization. In Proceedings of the 20th International Conference on Machine Learning. AAAI Press, pp.888–895,2003.

[2] I. Bose and A. C. M. Leung, "Unveiling the mask of phishing: Threats, preventive measures, and responsibilities," communications of the Association for Information Systems, vol. 19, pp. 544-566, 2007. [3] E. Kirda and C. Kruegel, "Protecting users against phishing attacks," The Computer Journal, 2005.

[4] E. Kirda and C. Kruegel, "Protecting users against phishing attacks," The Computer Journal, 2005.

[5] "Phishing activity trends report," Anti-Phishing Working Group, Tech. Rep., Jan. 2005. [Online]. Available: http://www. Antiphishing.org/reports/apwg report jan 2006.pdf

[6] B. Ross, C. Jackson, N. Miyake, D. Boneh, and J. Mitchell, "A browser plug-in solution to the unique password problem, http://crypto.stanford.edu/PwdHash/, 2005.

[7] J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, pp. 81–106, 1986.

[8] "Improved use of continuous attributes in c4.5," Artificial Intelligence Research, vol. 4, pp. 77–90, 1996.

[9] I. H. Witten and E. Frank, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.

[10] T. M. Mitchell, "Machine learning," 1997. [12] M. Dash and H. Liu, "Feature selection for classification," 1997.