# Introduction on Content Based Phishing Detection

**Shweta Dudhat[1] Priyank Patel[2] Nayan Mali[3] Romil Patel[4]**
[3,4]Assistant Professor
[1,2,3,4]SIE, vadodara

*Abstract*— Website phishing is the threatening challenge for the online society due to large number of transactions over the internet which happens on daily bases. Phishing tries to attempt to gather sensitive information by masquerading as a trustworthy entity in an electronic transaction/communication. The social networking sites like Facebook, Twitter and E-mails accounts are more affected from phishing or fake pages. The main idea behind writing this is to investigate the use of automated data mining ways in finding the complex problems of finding phishing websites for helping the users from being hacked. The approach for data mining is called Associative Classification method that suites best for finding phishing websites accurately. The common associative classification algorithm MCAC: "Multi-Label Classifiers based Associative Classification" to seek its applicability to the phishing. MCAC detects phishing websites with high accuracy than other algorithms and it generates hidden rules that other algorithms are unable to find and has improved predictive performance

***Key words:*** phishing detection, classification, data mining, security, MCAC algorithm

## I. INTRODUCTION

### A. *Phishing:*

Phishing is a typical classification problem in which the goal is to assign a test data one of the predefined classes [1]. A main security threat to online business comes from what becomes to be known as "phishing attacks". In such attacks malicious people create web pages that mimic the webpages of legitimate client of the legitimate site mistakenly access the faked web site and expose their financial and personal information to malicious people whom might use this information to perform illegal and criminal activities. There are many characteristics and indicators that can distinguish the original legitimate e-banking website from the phishing one. Phishing has a huge negative impact on organizations revenues, customer relationship, marketing efforts and overall corporate image. The aim of the phishing website is to steal the victims personal information by visiting and surfing a fake webpage that looks like a true one of a legitimate bank or company and asks the victim to enter personal information such as their username, account number, password, credit card number etc . Attackers might also commit identity theft crimes using victim's stolen information. it also damages the reputation of the attacked company

### B. *Two Most Popular Approaches in Designing Technical Anti Phishing Solution:*

*1) Blacklist Approach:*
where the requested URL is compared with a predefined phishing URLs.

*2) Search Approach:*
it is the based on search methods, where several website features are collected and used to identify the type of the website [2].

### C. *Associative Classification Works In Three Steps:*

- Step 1: discovering and generating the rules.
- Step 2: building the classifier.
- Step 3: prediction.

AC algorithms depend on two important thresholds: minimum support and minimum confidence. Minimum support represents the frequency of the attribute value and its related class in the training data set. Minimum confidence represents the frequency of the attribute value and it is related class in the training data set.\

### D. *Data Mining:*

Data mining is the process of searching through large amounts of data and picking out relevant information. It has been described as "the nontrivial extraction of implicit, previously unknown and potentially useful information from large dataset [3].

## II. LIFECYCLE OF PHISHING

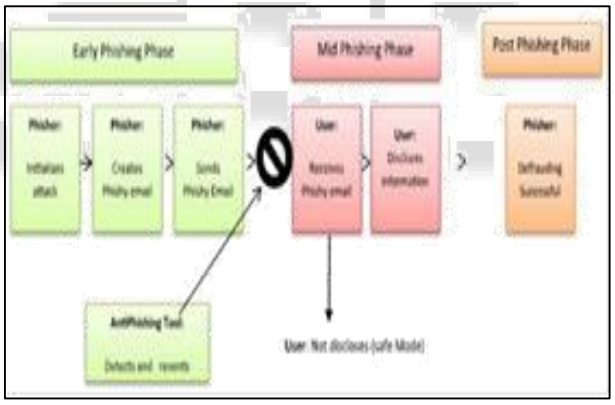


Fig. 1: Life cycle of Phishing

In Early Phishing phase, firstly Phisher creates Phishy email to initialize attack and sends to the User.

In Mid Phishing phase, user receives an email sent by the Phisher. Whenever user opens the email, the information will be disclosed to the Phisher. If user doesn't open the email, then the user is in Safe mode.

In Post Phishing phase, Phisher successfully gains the information of the user.

## III. WEB FEATURE RELATED TO PHISHING

a) IP address
b) Long URL
c) URL's having @ symbol
d) Prefix and suffix
e) Sub-domain (dots)
f) Misuse/fake of HTTPs protocol
g) Request URL
h) Server form handler

 

i) URL of anchor
j) Abnormal URL
k) Using pop-up window
l) Redirect page
m) DNS record
n) Hiding the links
o) Website traffic
p) Age of domain
rules of features:
1. IP address
Rule: if ip address exists in URL ? Phisy
Else =legit
2. Long URL
Rule : If URL length < 54 -> Legit
URL length => 54 and <= 75 -> Suspicious
Else -> Phishy
3. URL's having @ symbol
Rule :If URL has '@''-> Phishy
Else Legit
4. Prefix and suffix
Rule : If domain part has '-' -> Phishy
Else -> Legit
5. Sub-domain (dots)
Rule : if dots in domain < 3 ? Legit
Else if = 3 ?Susicious
Else -> Phishy
6. Misuse/fake of HTTPs protocol
Use of https & trusted issuer & age>= 2 years ->Legit using https & issuer is not trusted -> Suspicious
Else -> Phishy
7. Request URL
Rule : request URL < 22% ? legit
Request URL >= 22% and <=61% ? suspcious
Else ? phishy
8. Server form handler
Rule: SFH if 'about : blank' or empty ? Phishy
SHD redirects to different domain ? suspicious
Else ? legit
9. URL of anchor
Rule : URL Anchor % < 31% ? legit
URL Anchor % >= 31% and <=67% ? suspicious
Else ? Phishy
10. Abnormal URL
11. Using pop-up window
Rule: rightClick disabled ? phishy
rightClick showing alert ? suspicious
Else ? legit
12. Redirect page
Rule: redirect page #s >= 1 ? legit
Redirect page #s>1 and <=4 ? suspicious
Else ? Phishy
13. DNS record
Rule: no DNS record ? phishy
Else? legit
14. Hiding the links
Rule: change of statusbar onMouseOver ? Phishy
No change ? suspicious
Else?legit
15. Website traffic
rule: webTraffic < 150000 ? legit
Webtrafic >150000 ? suspicious
Else ? Phishy
16. Age of domain
Age of Domain : check on WHOIS databse
Rule: age <= 6 months ? legit
Else ? Phishy

## IV. STUDY

We have used the above given features to compare few algorithms such as Naïve Bayes, Update Table Naïve Bayes, Apriori algorithm. The result is as shown below.

### *A. Naïve Bayes:*

| TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|
| 0.904 | 0.05 | 0.936 | 0.904 | 0.919 | 0.981 | -1 |
| 0.95 | 0.096 | 0.926 | 0.95 | 0.938 | 0.981 | 1 |
| Weighted Avg. 0.93 | 0.076 | 0.93 | 0.93 | 0.93 | 0.981 | |

=== Confusion Matrix ===

| a | b | <-- classified as |
|---|---|---|
| 4427 | 471 | a = -1 |
| 305 | 5852 | b = 1 |

### *B. Simple Naïve Bayes:*

| TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|
| 0.904 | 0.05 | 0.936 | 0.904 | 0.919 | 0.981 | -1 |
| 0.95 | 0.096 | 0.926 | 0.95 | 0.938 | 0.981 | 1 |
| Weighted Avg. 0.93 | 0.076 | 0.93 | 0.93 | 0.93 | 0.981 | |

=== Confusion Matrix ===

| a | b | <-- classified as |
|---|---|---|
| 4427 | 471 | a = -1 |
| 305 | 5852 | b = 1 |

### *C. Apriori Algorithm:*

Minimum support: 0.85 (9397 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 3

Generated sets of large itemsets:

Size of set of large itemsets L(1): 12
Size of set of large itemsets L(2): 6
Size of set of large itemsets L(3): 2

Best rules found:

1) port=1 Iframe=1 9429 ==> RightClick=1 9423 conf:(1)
2) Iframe=1 10043 ==> RightClick=1 10035 conf:(1)
3) on_mouseover=1 Iframe=1 9529 ==> RightClick=1 9521 conf:(1)

4) port=1 9553 ==> RightClick=1 9512 conf:(1)
5) on_mouseover=1 9740 ==> RightClick=1 9665 conf:(0.99)
6) port=1 RightClick=1 9512 ==> Iframe=1 9423 conf:(0.99)
7) port=1 9553 ==> Iframe=1 9429 conf:(0.99)
8) port=1 9553 ==> RightClick=1 Iframe=1 9423 conf:(0.99)
9) on_mouseover=1 RightClick=1 9665 ==> Iframe=1 9521 conf:(0.99)
10) Shortining_Service=1 9611 ==> double_slash_redirecting=1 9422 conf: (0.98)

## V. CONCLUSION

We have employed Associative classification methods to classify URL as legitimate (ham) or phishing or suspicious URL. We have introduced some new type of features and implemented old features. In future we will use MCAC algorithm because of its high accuracy and high efficiency as compared to other algorithms.

## REFERENCES

[1] Neda Abdelhamid, Aladdin Ayesh, Fadi Thabtah – working on phishing detection based associative classification data mining ,2014.
[2] Aanchal Goel, Deepika Sharma- prevention from hacking attacks: phishing detection using associative classification data mining.
[3] Kantardzic and Mehmed. "data mining : concepts models, methods and algorithms"., John Wiley & sons.ISBN 0471228524. OCLC 50055336,2003.
[4] Maher Abumos, M.A. Hossain, Keshav Dahal, fadi Thabtah,"Associative classification techniques for predcting e-banking phishing websites",MCIT,IEEE,2010.
[5] mitesh dedakia, Khushali Mistry- working for phishing detection using content based ssociative classification data mining,july 2005.
[6] http://www.alexa.com (alexa the web information company-2011)
[7] Michael Kunz, Patrick Wilson, "Computer Crime and Computer Fraud", University of Maryland, 2004.

 