Impact Factor (SJIF): 3.632

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 3, Issue 5, May-2016

Content Based Phishing Detection

Shweta Dudhat [1], Priyank Patel [2], Nayan Mali [3]

[1] IT Dept, *SIE*
[2] IT Dept, *SIE*
[3] *Assistant Prof IT Dept, SIE*

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.***

Keywords-component - phishing detection, classification, data mining, security, MCAC algorithm

I. INTRODUCTION

The internet is not only important for individual users but also for organizations doing business online. Many of the organizations offer online trading and online sales of services and goods. Internet-users may be vulnerable to different types of online threat that may cause financial damages, identity theft, and loss of private information. Therefore, the internet suitability as a channel for commercial exchanges comes into question.

Phishing is a form of social engineering in which an attacker, also known as a phisher, attempts to fraudulently retrieve legitimate users' confidential or sensitive credentials by mimicking electronic communications from a trustworthy or public organization in an automated fashion. The word "phishing" appeared around 1995, when Internet scammers were using email lures to "fish" for passwords and financial information from the sea of Internet users; "ph" is a common hacker replacement of "f", which comes from the original form of hacking, "phraking" on telephone switches during 1960s. Early phishers copied the code from the AOL website and crafted pages that looked like they were a part of AOL, and sent spoofed emails or instant messages with a link to this fake web page, asking potential victims to reveal their passwords.

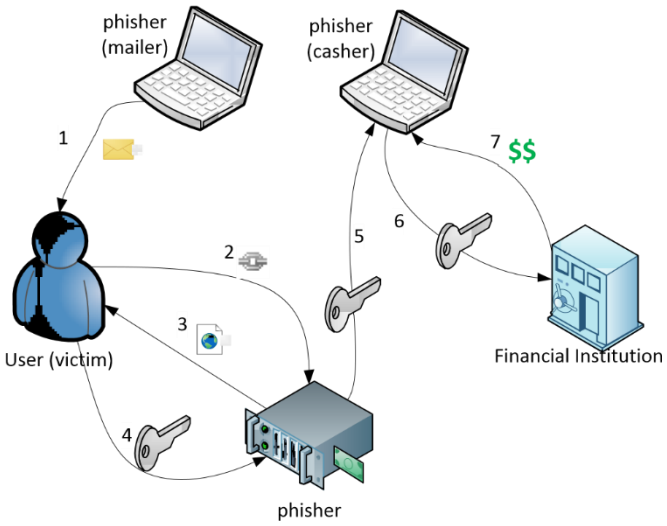


Figure 1: Phishing information flow

A complete phishing attack involves three roles of phishers. Firstly, mailers send out a large number of fraudulent emails (usually through botnets), which direct users to fraudulent websites. Secondly, collectors set up fraudulent websites (usually hosted on compromised machines), which actively prompt users to provide confidential information. Finally, cashers use the confidential information to achieve a pay-out. Monetary exchanges often occur between those phishers.

II. Related Work

The approach described here is to apply data mining algorithms to assess phishing website risk on the 30 characteristics and factors which stamp the forged website. We utilized data mining classification and association rule approaches in our new phishing website detection model to find significant patterns of phishing characteristic or factors in the phishing website archive data. Particularly, we used a number of different existing data mining association and classification techniques.

Feature set	Phishing feature indicator
Domain identity and URL	via IP address Require URL URL of anchor DNS details Strange URL
Encryption and security	SSL certificate Certification authority Strange cookie Distinguished names certificate (DN)
Java script and source code	Redirect pages Straddling attack Pharming attack Using onMouseOver Server form handler
Contents and page style	Spelling mistake Replicating a website ‘‘Submit’’ button Via pop-up windows Disabling right-click
Web address bar	Long URL address Replacing similar characters for URL Adding prefix or suffix Using the ‘@’ to confuse Using hexadecimal character codes
Social human factor	Much stress on security and response Generic welcome Buying time to log on accounts

Table 1: Phishing Indicators and their criteria

1. The approach described here is to apply data mining algorithms to assess website phishing risk on the Existing & New characteristics and factors which stamp the forged website.

2. Associative and classification algorithms can be very useful in predicting Phishing websites.

3. It can give us answers about what are the most important phishing website characteristics and indicators and how they relate with each other.

4. The choice of MCAC algorithm is based on the fact that it combines both approaches to generate a set of rules.

5. Associative classifiers produce more accurate classification models and rules than traditional classification algorithms.

III. Problem Statement

Multi label Class Associative Classification (MCAC) is a lasted technique to detecting phishing using various features of website. This method generate new hidden rule which cannot be generate by any other method. Its improve classifier predictive performance. There is one major limitation of this method is that; this method does not consider a content based feature to detecting phishing activity.

IV. Proposed Work

In existing MCAC algorithm it was consider 16 different features of website from URL and Domain identity, Security and Encapsulation, source code and JavaScript etc to detecting the behavior of website[1].

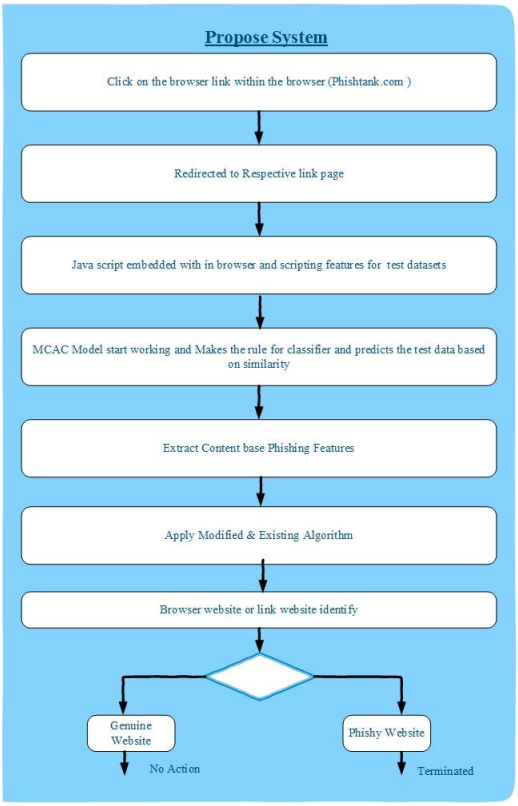


Figure 2: Proposed work

Fig. 2 shows a new modified proposed work for this research work. In that flow firstly users click on the browser link. If user click than it is redirected respective page. After redirected on website, Java script embedded within browser and extracts the features of websites and makes it suitable for test data. After that MCAC algorithm start working and it is makes association rule for extracted features. After that based on that rule it is classify the websites in terms of legitimate and phishing.

This work is done in MCAC existing methods.

Now Proposed work flow also perform the above step and add 2 steps in that which are based Java script we extract content based features of website. And applied modified algorithm or existing algorithm on this features and will make association rule and classify website after applied content based filtering algorithm.

Proposed Content Based Features

In this section, it includes some proposed content and page style based features of website which will considered in research work.

a) Spelling Error
b) Copying Website
c) Using Forms with Submit Button
d) Disabling Right-Click
e) Using Pop-Ups Windows

1. Spelling Error

Rule : if Spelling error is their -> Phishy
else -> legit

2. Copy website
Rule : if path does not match -> phishy
else -> legit

3. Using Forms with submit button
Rule : if forms with submit button >= 1 -> phishy
else -> legit

4. Disabling Rightclick
Rule : if check disable rightclick = yes -> phishy
else check disable rightclick = no -> legit

5. Using Pop-Ups Windows
Rule : if Using Pop-Ups Windows >= 2 and <= 4 ? legit
Using Pop-Ups Windows > 4 ? suspcious
Else Using Pop-Ups Windows =1 ? phishy

V. Algorithm

In this section algorithm is given. This algorithm is used MCAC algorithm and this is a algorithm steps of the phishing detection system.

Input: Training Dataset (D), Features Criteria (as minimum support minSup and minimum confidence minConf

Output: Classifier (3 Different Class)
Pre-processing: Change URL pattern format(if required)
Step 1
a) Input dataset value
b) Extract following Features of websites
c) URL & Domain Identity
d) Security & Encryption
e) Source Code & Java script
f) web address

Step 2
a) Extract the Content Base features of websites

Step 3
a) Generate association rule for frequent value
b) Convert frequent value that passes minConf and minSup to single label rule
c) Merge any two single label rule to derive multi label rules

Step 4
a) Sort the rule classify the test data

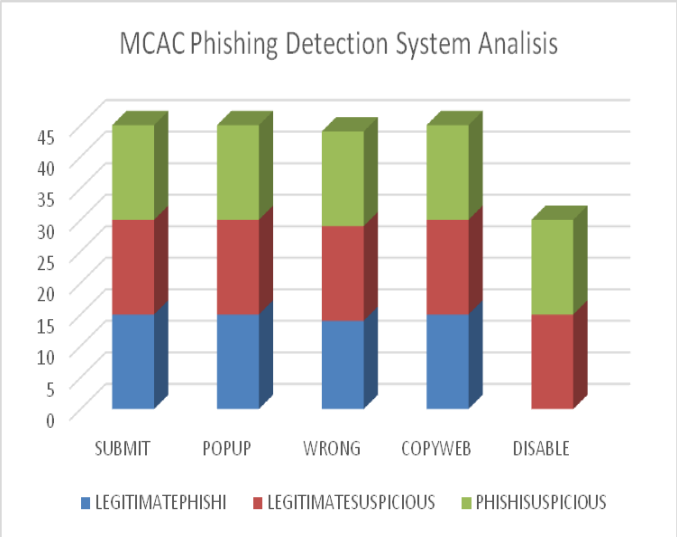
Experimental Analysis

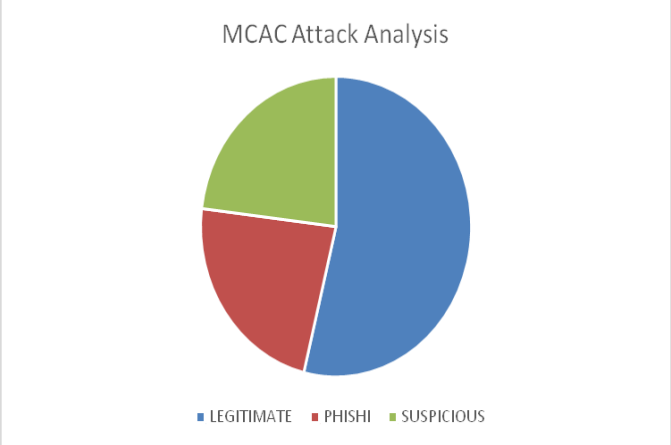
Database Description
In this section, we are discuss detail of database. We are using two publicly freely available datasets are used to make feasible our experimental environment of our proposed method.

We used PhishTank dataset. PhishTank is a free community site where anyone can submit, verify, track and share phishing data. This dataset is in the form of .csv file format.

VI. Experimental Result

Result content base on recent phishtank dataset some links.





Conclusion

In this research work, develop a modified system that consider content based features of website such as spelling error, coping website, using forms with submit button, disabling right click, using pop- up windows. In this research work we are experiment existing method and proposed work for single URL and also on dataset and on live dataset. In this research work, it is also listed comparison of existing and proposed work. In near future it is possible to add more content base features to get more accuracy. It is also possible to use content base algorithm instead of MCAC algorithm to design more accurate and get higher security.

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]. http://www.alexa.com (alexa the web information company-2011)
[6]. Michael Kunz, Patrick Wilson, “Computer Crime and Computer Fraud”, University of Maryland,2004.