PDS - Phishing Detection Systems

Ria Sankhyan[1], Ankit Shetty[2], Lubdha Dhanopia[3], Chetan Kaspale[4], Prof. Gayatri Dantal[5]

[1,2,3,4,5] *Department of Information Technology, Shree L.R. Tiwari college of Engineering, Maharashtra, India*

---***---

Abstract - *There is abundant data available online. However, the data is prone to various malicious activities or the owner can be spoofed. This event is known as phishing. Phishing is an criminal act that takes place when a malicious webpage impersonates as true webpage in order to acquire sensitive information from the user. However, there are tools to identify phishing sites but they are generally in data. Phishing attack continues to pose a serious threat for users and bothersome risks. This project focuses on discerning the significant features that distinguish between legitimate and phishing URLs. The rules obtained are interpreted to accentuate the features that are more prevalent in phishing URLs. Classification is an important problem in data mining. A number of popular classifiers build decision trees to spawn class models. The data classification is based on ID3 Decision Tree algorithm which result in accuracy, the data is estimated using entropy based cross validations and partition techniques and the results are compared thereby identifying the URL.*

***Key Words*: Phishing, prediction, machine-learning, data mining, ID3.**

1. INTRODUCTION

Living in the space age technology, use of internet is mandatory, even for the smallest of tasks. With the increased use of the internet, web services and growth of the Internet environment web attacks have increased in quantity and advanced in quality. Not only the attacks but the number of spam emails rapidly increasing, currently estimated that more than half of all emails are spam — the number of spam emails containing malicious attachments is rising. Phishing, a social engineering attack that targets user's private information is a complete threat to the securing system used nowadays. According to research of the AntiPhishing Working Group (APWG), 85,062 phishing sites were globally detected in the second quarter of 2010. Annual damage instigated by phishing was measured at $5.9 billion.

2. LITERATURE REVIEW

Here we will elaborate the aspects like the literature survey of the project and what all projects are existing and been actually used in the market which the makers of this project took the inspiration from and thus decided to go ahead with the project covering with the problem statement.

In this paper, authors Routhu Srinivasa Rao and Alwyn Roshan Pais [1], have proposed a blacklisting method, they've enhance over the previous blacklisting techniques by using key discriminatory features, and extracted them from the source code of the website for detection pf phishing attacks. They've made use of Simhash algorithm, by means of which, they've achieved 84.36% of detection accuracy.

The authors Choon Lin Tan, Kang Leng Chiew, KokSheik Wong and San Nah Sze [2], have proposed a detection technique which detects the difference between target and actual identities of webpage. Their experimental results revealed the system to have a better performance than existing conventional phishing detection methods considered.

in this survey paper [3], authors Clifton Phua, Vincent Lee, Kate Smith and Ross Gayler, have surveyed the various papers on fraud detection techniques that use data mining, within last ten years and have tried to enhance the knowledge on various terms like fraudsters, fraud etc. as well as covered as much technical aspect as possible.

3. SYSTEM ARCHITECTURE.

A set of data is taken which has both phishing and non-phishing site URLs.

In the first phase, a classifier is produced using URLs of phishing sites and legitimate sites collected in advance. the URLs are then uploaded to the feature extractor wherein the values from pre-defined URL-based features which are implemented are extracted. The features extracted are stored as input are then passed to the classifier generator.

The classifier generator is a classifier which compares the features to the entered url. This is done in the detection phase, to check whether the requested site is a phishing site or not. When a request for the authenticity of the page occurs, the URLs is transferred to the feature extractor, wherein the features are entered into the classifier. Then after comparison the classifier will determine whether the requested site is legitimate site or not based on the educated information.

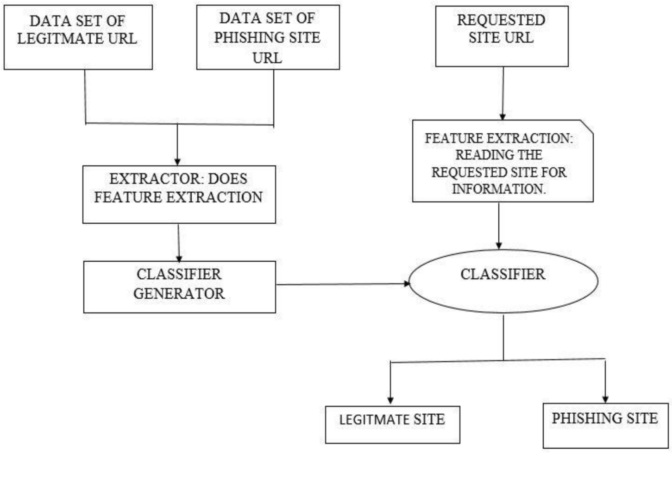

DATA SET OF LEGITMATE URL
DATA SET OF PHISHING SITE URL
REQUESTED SITE URL
FEATURE EXTRACTION: READING THE REQUESTED SITE FOR INFORMATION.
EXTRACTOR: DOES FEATURE EXTRACTION
CLASSIFIER GENERATOR
CLASSIFIER
LEGITMATE SITE
PHISHING SITE

Fig -1: Block Diagram for Phishing Detection System.

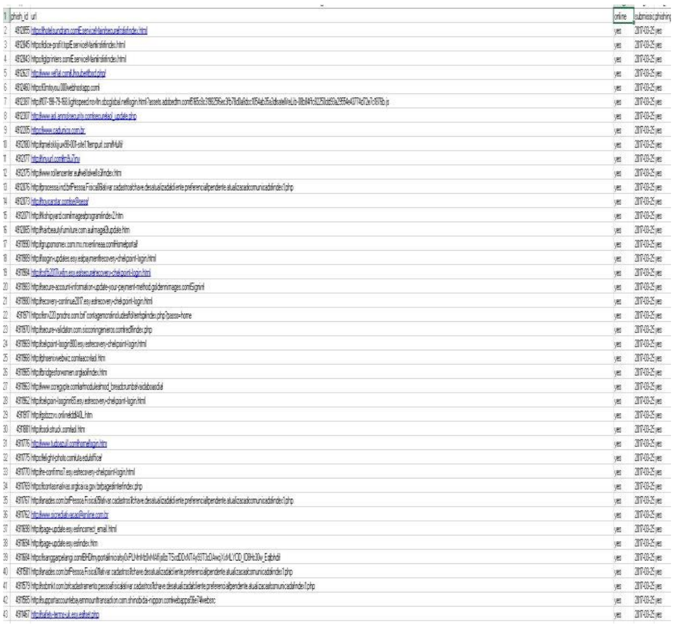
- Safety of Google
- Suggestion (path domain) whether it is present in whitelist.
- Google page rank PageRank value of domain
- Alexa rank /Alexa Rank value of domain
- Alexa reputation Via IP address
- Length of URL
- Suspicious character, etc.

4. IMPLEMENTATION

To start the training we first prepare a dataset. This dataset is then uploaded into the classifier and it is then sent for feature extraction. Following are the steps:

a. Data Preparation.
b. Feature Extraction.
c. Implementation of ID3 algorithm.
d. Model.

Step 1:Data Preparation: The data set obtained in this study was obtained from Phishtank a website which contains data about various illegitimate websites.



	phish_id	url	online	submiss	phishing
2	[illegible]	[illegible]	yes	2017-03-25	yes
3	[illegible]	[illegible]	yes	2017-03-25	yes
4	[illegible]	[illegible]	yes	2017-03-25	yes
5	[illegible]	[illegible]	yes	2017-03-25	yes
6	[illegible]	[illegible]	yes	2017-03-25	yes
7	[illegible]	[illegible]	yes	2017-03-25	yes
8	[illegible]	[illegible]	yes	2017-03-25	yes
9	[illegible]	[illegible]	yes	2017-03-25	yes
10	[illegible]	[illegible]	yes	2017-03-25	yes
11	[illegible]	[illegible]	yes	2017-03-25	yes
12	[illegible]	[illegible]	yes	2017-03-25	yes
13	[illegible]	[illegible]	yes	2017-03-25	yes
14	[illegible]	[illegible]	yes	2017-03-25	yes
15	[illegible]	[illegible]	yes	2017-03-25	yes
16	[illegible]	[illegible]	yes	2017-03-25	yes
17	[illegible]	[illegible]	yes	2017-03-25	yes
18	[illegible]	[illegible]	yes	2017-03-25	yes
19	[illegible]	[illegible]	yes	2017-03-25	yes
20	[illegible]	[illegible]	yes	2017-03-25	yes
21	[illegible]	[illegible]	yes	2017-03-25	yes
22	[illegible]	[illegible]	yes	2017-03-25	yes
23	[illegible]	[illegible]	yes	2017-03-25	yes
24	[illegible]	[illegible]	yes	2017-03-25	yes
25	[illegible]	[illegible]	yes	2017-03-25	yes
26	[illegible]	[illegible]	yes	2017-03-25	yes
27	[illegible]	[illegible]	yes	2017-03-25	yes
28	[illegible]	[illegible]	yes	2017-03-25	yes
29	[illegible]	[illegible]	yes	2017-03-25	yes
30	[illegible]	[illegible]	yes	2017-03-25	yes
31	[illegible]	[illegible]	yes	2017-03-25	yes
32	[illegible]	[illegible]	yes	2017-03-25	yes
33	[illegible]	[illegible]	yes	2017-03-25	yes
34	[illegible]	[illegible]	yes	2017-03-25	yes
35	[illegible]	[illegible]	yes	2017-03-25	yes
36	[illegible]	[illegible]	yes	2017-03-25	yes
37	[illegible]	[illegible]	yes	2017-03-25	yes
38	[illegible]	[illegible]	yes	2017-03-25	yes
39	[illegible]	[illegible]	yes	2017-03-25	yes
40	[illegible]	[illegible]	yes	2017-03-25	yes
41	[illegible]	[illegible]	yes	2017-03-25	yes
42	[illegible]	[illegible]	yes	2017-03-25	yes
43	[illegible]	[illegible]	yes	2017-03-25	yes

Fig -2: Data Taken

Step 2: Feature Extraction: Feature extraction is the process to extract the features from the URLs to facilitate subsequent learning and generalization step to acquire better interpretation

Step 3: Implementation of ID3 algorithm: The ID3 algorithm is used in this project, for training model. When the algorithm is applied, it generates a rule set based on the observed pattern of data. On the basis of this rule set, the system training is done and the model is created.

ID3 algorithm:

The project makes use of ID3 algorithm. ID3 stands for Iterative Dichotomiser 3. The algorithm was developed by Ross Quinlan and is used to generate a decision tree from a given dataset. ID3 works mainly on three things, firstly the entropy of each attribute, second information gain and third, entropy of whole dataset, using these three, it picks a root node. The condition for selection of root node is that, the attribute with lower value of entropy (or higher value of information gain.) becomes the root node. This carries on until the last element of data that can provide some substantial information is not used.

- Calculate the entropy of every attribute
- Then it is split into subsets
- Make a decision tree node containing that attribute
- Recurse on subsets using remaining attributes.

Step 4: Model: Once the system training is done, our general model will be ready for using.

5. RESULTS:

The results for Detection of Phishing Sites Using URL Features are shown below.

Step 1: System Training

The user trains system through this module.

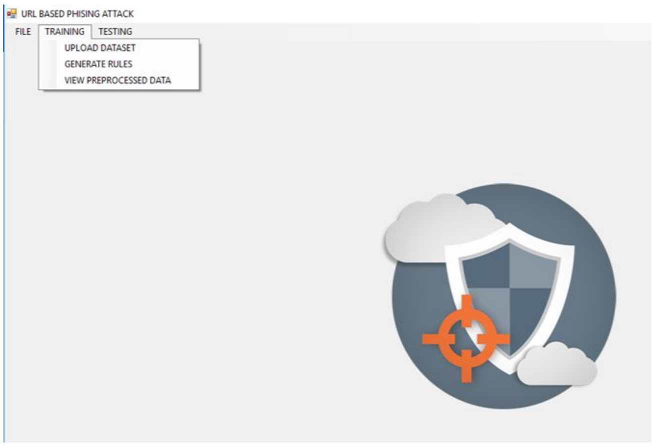
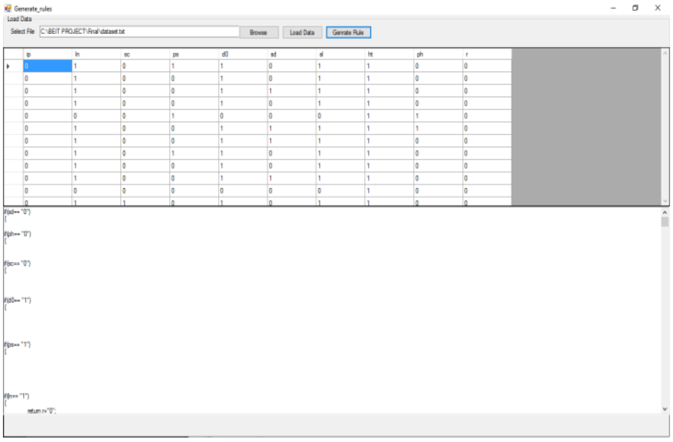


Fig- 3: System Training

Step 2: Generated Rule Set

Once the data is fed to the system, the rule set is generated based on which the model is created.



Generate_rules

Load Data

Select File: C:\BEIT PROJECT\Final\dataset.txt

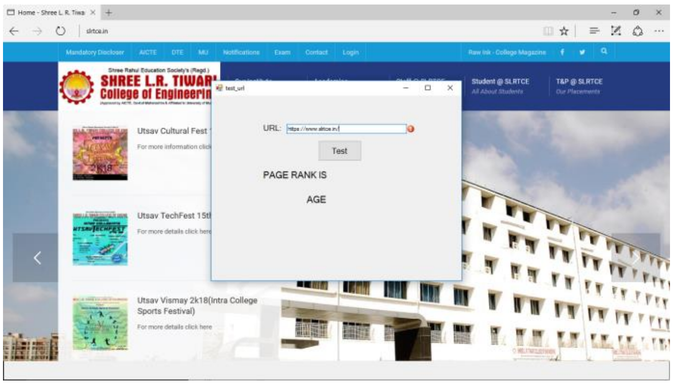
ip	ln	sc	ps	d0	sd	sl	ht	ph	r
0	1	0	1	1	0	1	1	0	0
0	1	0	0	1	0	1	1	0	0
0	1	0	0	1	1	1	1	0	0
0	1	0	0	1	0	1	1	0	0
0	0	0	1	0	0	0	1	1	0
0	1	0	0	1	1	1	1	1	0
0	1	0	0	1	1	1	1	0	0
0	1	0	1	1	0	1	1	0	0
0	1	0	0	1	0	1	1	0	0
0	1	0	0	1	1	1	1	0	0
0	0	0	0	0	0	0	1	0	0
0	1	1	0	1	0	1	1	0	0

```
if(sd== "0")
{

if(ph== "0")
{

if(sc== "0")
{

if(d0== "1")
{

if(ps== "1")
{

if(ln== "1")
{
        return r="0";
```

Fig- 4: Generated Rule Set

Step 3: Giving Input

The user gives the input to the system through this module, it is this input that will be checked by the system for potential danger.



test_url

URL: https://www.slrtce.in/

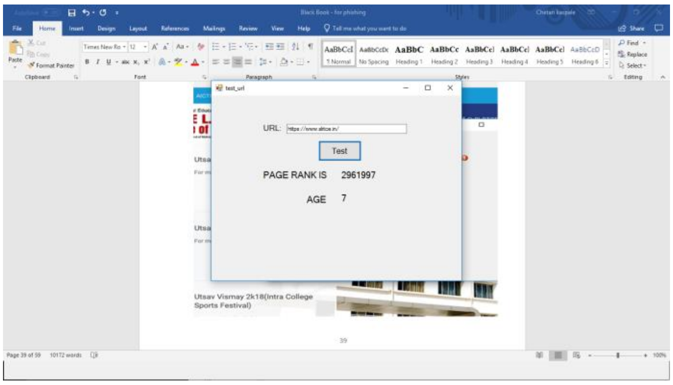
PAGE RANK IS

AGE

Fig- 4: Taking Input

Step 4: Outcome

In this step, the system provides with the result of whether or not the URL is safe, and also provides with the Alexa page rank and age of the website.



test_url

URL: https://www.slrtce.in/

PAGE RANK IS 2961997

AGE 7

Fig- 5: Final result

6. CONCLUSIONS:

In future work, we intend to address the time-intensive disadvantage of the phishing attack technique. With a large number of features, it is time-consuming for the URL based phishing attack approach to generate classifiers and perform classification. Therefore, we will apply algorithms to reduce the number of features and thereby improve performance. In addition, we will examine a new phishing detection technique that uses URL-based features, and with working of ID3 algorithm and features phishing website can be identified. Furthermore, a bulk of websites can be tested in the further process. The expected accuracy rate for this system is for detecting phishing websites is calculated to be 60%-90%. Users just need to provide the URL of the website whose legitimacy needs to be determined.

REFERENCES

[1] Routhu Srinivasa Rao, Alwyn Roshan Pais, "An Enhanced Blacklist Method to Detect Phishing Websites", 2017, ICISS, pp 323-333

[2] Choon Lin Tan, Kang Leng Chiew, KokSheik Wong, San Nah Sze, "PhishWHO: Phishing webpage detection via identity keywords extraction and target domain name finder", 2016 August, Volume 88, Decision Support Systems, pp 18-27

[3] Clifton Phua, Vincent Lee, Kate Smith, Ross Gayler," A Comprehensive Survey of Data Mining-based Fraud Detection Research"pages.