Fraud Website Detection using Data Mining

Urvashi Prajapati, Neha Sangal
Department of Information Technology,
K. J. Somaiya College of Engineering,
Mumbai, India

Deepti Patole
Department of Information Technology,
K. J. Somaiya College of Engineering,
Mumbai, India

ABSTRACT
Phishing attack is used to steal confidential information of user. Fraud websites appear similar to genuine websites with the logo and graphics of trusted website. Fraud Website Detection application aims to detect fraud websites using data mining techniques. This project provides intelligent solution to phishing attack. W3C standard defines characteristics which can be used to distinguish fraud and legal website. This application extracts some characteristics from URL and source code of a website. These features are used for classification. RIPPER algorithm is used to classify the websites. After classifying the websites, the application sends notification email to the administrator using WHOIS protocol. The administrator may block the fraud website after verification.

Keywords
Phishing, JRip, RIPPER, Fraud website, source code, data mining, WHOIS.

1. INTRODUCTION
Phishers have many tactics and approaches to perform a well-sophisticated phishing attack. Fraud websites are used to steal confidential information such as passwords, usernames, security codes, credit card numbers, etc [1] .Fraud websites appear like genuine websites, even to the point of using the logos and references same as that of legitimate website. Usually people trust the information they receive from websites and enter their confidential information.

According to Anti-Phishing Working Group (APWG), there were 47,094 unique phishing websites detected during 4th Quarter 2014. There were 437 brands targeted by phishers in Q4 [2].The different online industries like financial, payment services, social networking, government, email services etc. are affected by phishing.

There are two categories for Fraud Website detection:

List based and Heuristic based [3]. List based detection is done with the help of black-list and white-list. Blacklist consists of phishing URLs. Whitelist holds legitimate URLs.

When user tries to search for a particular website the browser queries the blacklist. If the website is found in the blacklist, then the user is denied access. The advantages of list-based approach are speed and simplicity. The drawback of this approach is that it takes time to add phishing site to a blacklist once it is detected.

Heuristic-based approach checks one or more characteristics like URL, HTML source code or the page content. This approach uses data mining algorithms to classify legal and fraud websites. The strength of this approach is its ability to detect zero-day phishing attacks.

Fraud website detection application uses heuristic-based approach and corrective measure against phishing attack.

2. RELATED WORK

2.1 Algorithm
Classification models predict categorical class labels. The repeated incremental pruning to produce error reduction (ripper) is a direct classification algorithm which extracts rules directly from datasets. Ripper aims at increasing the accuracy of rules by replacing or revising individual rules. In ripper, rules are learned incrementally. One rule can cover more than one attribute of a class. The algorithm chooses one of the classes as positive class, as well as the other as negative class, for 2-class problem,. Firstly, the algorithm learns rules for positive class and negative class is default class. Rule generation and rule optimization is carried out. Ripper gives fast and efficient results than decision tree while dealing with large datasets.

The algorithm is briefly described as follows [4]:

First step is to initialize RS= {} and follow the steps given below for each class ranging from less conventional to more frequent one.

DO:

1. Building stage:

Repeat step 1.1 and step 1.2, stop when the Description Length (DL) is 64 bits greater than the smallest DL obtain till now, or there are no positive examples, or error rate >= 50%.

1.1. Grow phase:

Rules are increased by greedily adding conditions to the rule until the rule becomes more perfect (i.e. 100% accurate).[5] Every possible value of each attribute is evaluated by the procedure and the condition with the highest Information Gain is selected. Here,

$$Information\ Gain = p\ (\log(p/t) - \log(P/T)$$

Where *p: number of instances covered by rules that are positive*

t: total instances covered by rule

P: positive numbers before the new condition was added

T: total numbers before the new condition was added

1.2. Prune phase:

After rules are induced, performance of rule is tested by Pruning Metric. Here,

$$Prunning\ Metric = \frac{p}{p+n}$$

Where *p: total number of positive instances*

n: total number of negative instances

It allows the pruning of final sequence of conditions (or antecedents) as well as shows the accuracy of rules [6]. Conjunct is removed from the rule only if the metric is improved after pruning.

2. Optimization stage:

Using procedure 1.1 and 1.2, generate and prune two variants of each rule Ri. Empty rule generates one variant. Whereas greedily adding antecedents to the original rule resulted in generation of other variant.

Then the computation of original rule and the smallest possible DL for each variants is done. The variant having the minimal DL is selected as the final representative of Ri in the ruleset. After all the rules in {Ri} have been examined and if there are still residual positives, more rules are generated based on the residual positives using Building Stage again.

3. Delete stage:

The rules that increase the DL of the ruleset then delete that rules. Add resultant ruleset to RS.

ENDDO

2.2 Properties of Phishing Attacks

Some of the properties of phishing attacks are [5]:

- **Short lived:** The duration of fraud websites is very less as compared to legal websites. It may be live for just few hours or days.
- **User Input:** Most of the fraud websites contain web forms asking user for confidential information such as credit card details, password, etc.
- **Mimicry:** Most of the fraud websites look similar to legal websites. Phishers link their website's images and logos to legal website domain.

2.3 Features

Some features that are used to distinguish fraud websites from legitimate ones are as follows [7]:

- **Using '@' Symbol**

If URL contains "@" symbol, then while reading an internet address, the web browser ignores everything preceding the "@" symbol. Therefore, http://www.flipkart.com@fraud.com would be "fraud.com".

- **IP Address in URL**

IP addresses are used to uniquely identify a host machine in a network. Sometimes, legal websites also use IP Address for internal private devices. Phishers use IP address in URL to hide the domain name of the website, e.g. http://172.45.3.256. Phishers may also use IP address with legal URL or keywords such as "http://www.paypal.com@45.35.82.216". User may feel that he is accessing PayPal website but in reality he is navigated to 45.35.82.216.

- **Iframe**

It is a html tag and is used to embed another document within the current HTML document. Phishers use borderless iframe and inject malicious code into it. Phishers may also insert a web form using iframe, asking for user details. User feels that he is on a trusted website and may enter confidential information.

For example, <iframe src='http://www.fraudwebsite.com' FRAMEBORDER='0' width='500' height='340' scrolling='auto'></iframe>.

- **Image**

Usually, phishers use logos from the legitimate target page. All the images in the website should belong to the same domain. If the images have been linked to another website, then it is considered as a phishing character.

- **Redirect**

Redirection is used to navigate from one URL to other. It can be used to redirect to malicious website. For example, a link, http://www.abc.com/login.php?redirect=http://www.abc.com/ home.phpredirects to http://www.abc.com/home.php

- **Submit**

Usually, action attribute of form in fraud websites contains an email id or refers to a different domain [8]. For example, <form action=abc@pqr.com target="top">.

- **Hexadecimal characters**

Hexadecimal characters preceded with '%' symbol can be used in URL of a website. Browsers can interpret hexadecimal codes. Hexadecimal values can be used to hide malicious URLs. For example,

Dotted Quad Notation : 192.168.1.1

Hexadecimal Format : 0xc0a80101

2.4 WHOIS

WHOIS is an internet program which allows user to query a database [9].WHOIS database stores the information about registered users, domain name, and IP address block. The WHOIS protocol also stores as well as delivers content of database in a human-readable format.

WHOIS information is useful to inform about fraudulently registered domain names the victims of prior identity theft through name, email address and contact numbers. This allows respective registrars to take action on domains that are part of current or future phishing attacks as early as possible. [10]

3. EXPERIMENTAL WORK

The fraud website detection system has been implemented using Netbeans 8.0.2 IDE. JFrame was used to create interface of this application. Figure 1 shows the work flow of Fraud Website Detection.

3.1 Data Collection

Dataset for the application consist of legitimate as well as fraud websites. Yahoo directory is crawler based web directory, includes all trusted websites. Legitimate websites were collected from Yahoo Directory [11] and DMOZ [12]. Phishtank database includes URLs of a fraud websites along with a screenshot, time of report and status of a website. So, fraud websites are collected from Phishtank[13]. Data is collected in .csv file. This .csv file is then used for extracting features from URL of the website. The aim is to identify strategies that were used by hackers and to gather trends used for different phishing attacks techniques.

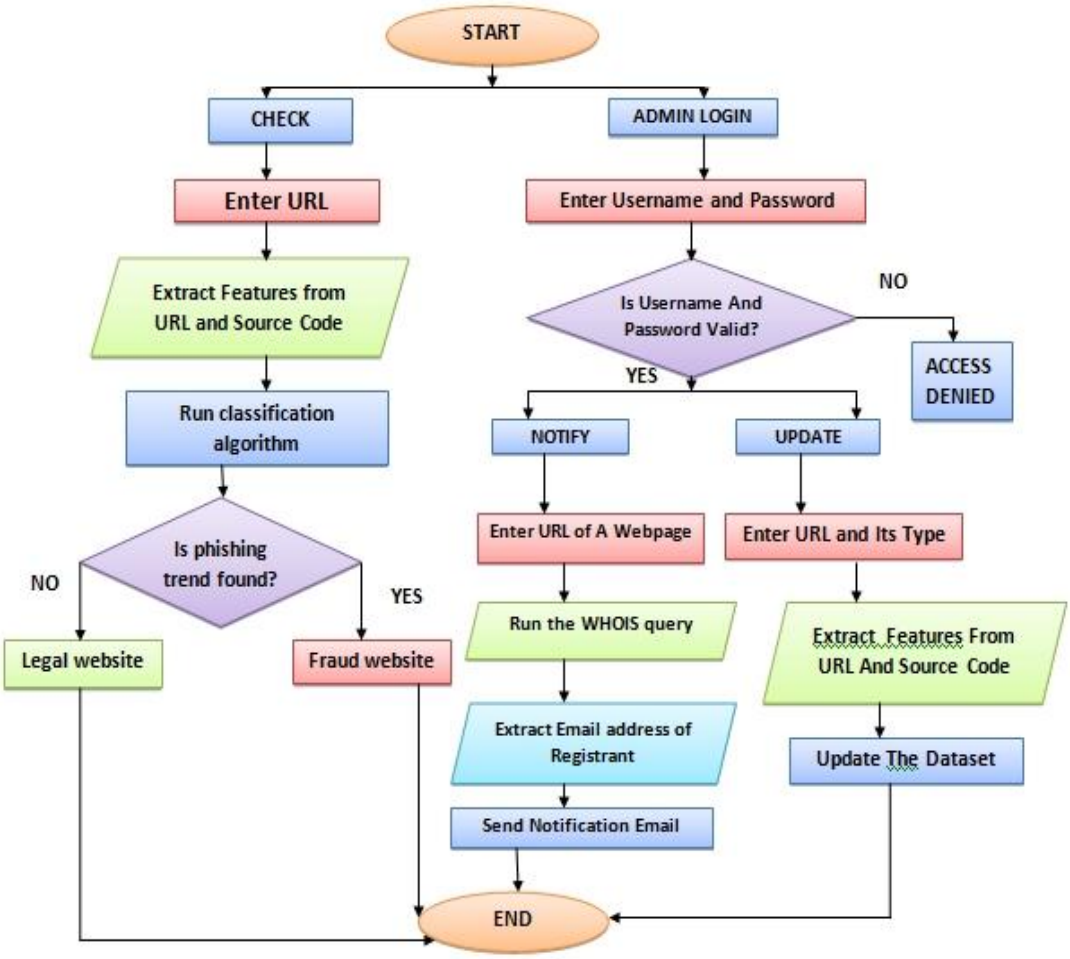

Figure 1 Fraud Website Detection Flow Chart

3.2 Features Extraction

A total of 23 features are extracted. Both lexical and binary features are considered. Features like length of URL, host length, special characters and hexadecimal characters are taken from URL. Iframe and image URL are extracted from web page source code. A total of 1250 URLs are considered for statistics where count of legal and fraud websites is 625 each. Table 1 and Table 2 show Lexical Features as well as Binary Features statistics respectively.

Table 1 Lexical Features Statistics

Features		Min	Max	Median
No. of dots	Legal	1	5	2
	Fraud	1	28	2
Length of URL	Legal	17	104	29
	Fraud	18	504	62
Host length	Legal	9	37	18
	Fraud	5	109	19
Special characters	Legal	2	14	3
	Fraud	2	67	6

Table 2 Binary Features Statistics

Features		Count
Presence of Hexadecimal characters in URL	Legal	34
	Fraud	217
Presence of IP address in URL	Legal	0
	Fraud	11
Presence of image with different website's URL in source code	Legal	522
	Fraud	618

3.3 Algorithm Implementation

RIPPER algorithm is used for classification. Netbeans is used for this purpose. Weka.jar files are used for implementing the algorithm. Figure 2 shows the results of implementation of RIPPER. 1250 URLs are given as input for training set. The algorithm is tested on 250 URLs containing 125 legal and 125 fraud URLs. The accuracy obtained is 86.4%.Table 3 shows the confusion matrix of testing dataset.

Table 3 Confusion Matrix

Actual \ Predicted	Legal	Fraud
Legal	103	22
Fraud	12	113

Figure 2 Result of Implementation of RIPPER

3.4 Notification

WHOIS information is the most important tool. It is used to locate and communicate with service providers, registrant and site owners. When administrator found a website to be fraud, system will run WHOIS query on URL of the website. In case of domain names that are fraudulently registered, this system will send notification email which may cause removal of webpage. Figure 3 shows result of notification module.

Figure 3 Notification Module

4. CONCLUSION

Nowadays, it is crucial to detect fraud website on zero day as the fraud websites are short lived, and are designed to create maximum damage before getting tagged and listed as black listed website. List based detection and Heuristic-based detection are two approaches for detecting fraud websites. List based detection is done with the help of black-list and white-list. This approach is unable to detect fraud websites on zero day or before the fraud website is blacklisted. Use of heuristic-based detection approach in Fraud website detection application, enables it to detect fraud websites before they are blacklisted. The application checks one or more characteristics like URL, HTML source code or the page content. The classification module of the application, which consists of data Mining Algorithm 'RIPPER', provides classification of any given website as Fraud or Legal. The application also takes corrective measure against Fraud Website by reporting about the high possibility of the website in question, being fraud to respective authority. Thus, the application will prove to be useful to reduce the risk of phishing attack by preventing users from entering confidential information in fraud websites.

5. FUTURE SCOPE

The application can be developed as a plug-in for web browser. This will warn the user regarding fraud website in real time while browsing the internet. So, the application will become more user-friendly.

6. ACKNOWLEDGMENT

Ms. Urvashi Prajapati and Ms. Neha Sangal would like to thank Prof. Deepti Patole for her support and guidance throughout the course of the project and would like to acknowledge K. J. Somaiya College of Engineering, for the valuable information provided by them in their respective fields.

7. REFERENCES

[1] Peter Stavroulakis, Mark Stamp,"Handbook of Information and Communication Security",Springer.

[2] Phishing statistics, https://docs.apwg.org/reports/apwg_trends_report_q4_2014.pdf

[3] M. Dunlop, S. Groat, and D. Shelly," GoldPhish: Using Images for Content-Based Phishing Analysis", in the Fifth International Conference on Internet Monitoring and Protection, 2010.

[4] JRip algorithm pseudocode, http://weka.sourceforge.net/doc.dev/weka/classifiers/rules/JRip.html

[5] Robert Stahlbock, Sven F. Crone, Stefan Lessmann ,"Data Mining: Special Issue in Annals of Information Systems",Springer.

[6] Zhongyu Lu, "Information Retrieval Methods for Multidisciplinary Applications", Information Science Reference.

[7] Mohammad, R., Thabtah, F., & McCluskey, L. (2012). — An assessment of features related to phishing websites using an automated techniquel. In The 7th international conference for internet technology and secured transactions(ICITST-2012). London: ICITST.

[8] Omar Abdullah Batarfi Mona Ghotaish Alkhozae, "Phishing websites detection based on phishing

characteristics in the webpage source code," International Journal of Information and Communication Technology Research, October 2011.

[9] Garth O. Bruen ,"WHOIS Running the Internet: Protocol, Policy, and Privacy",Wiley.

[10] Ihab Shraim, Laura Mather, Patrick Cain, Rod Rasmussen, "Advisory on Utilization of Whois Data For Phishing Site Take Down ", APWG Internet Protocol Committee,March 2008.

[11] Data for legal websites, http://www.dmoz.org/

[12] Data for Fraud websites, https://www.phishtank.com/