# Intelligent Rule based Phishing Websites Classification

Rami M. Mohammad
School of Computing and Engineering
University of Huddersfield
Huddersfield, UK
rami.mohammad@hud.ac.uk

Fadi Thabtah
School of MIS
Philadelphia University
Amman, Jordan
ffayez@philadelphia.edu.jo

Lee McCluskey
School of Computing and Engineering
University of Huddersfield
Huddersfield, UK
t.l.mccluskey@hud.ac.uk

*Abstract* — **Phishing is described as the art of emulating a website of a creditable firm intending to grab user's private information such as usernames, passwords and social security number. Phishing websites comprise a variety of cues within its content-parts as well as browser-based security indicators. Several solutions have been proposed to tackle phishing. Nevertheless, there is no single magic bullet that can solve this threat radically. One of the promising techniques that can be used in predicting phishing attacks is based on data mining. Particularly the "induction of classification rules", since anti-phishing solutions aim to predict the website type accurately and these exactly fit the classification data mining. In this paper, we shed light on the important features that distinguish phishing websites from legitimate ones and assess how rule-based classification data mining techniques are applicable in predicting phishing websites. We also experimentally show the ideal rule based classification technique for detecting phishing.**

***Keywords- Website features*, *Phishing, Security, Rule based Classification, Data Mining.***

## I. INTRODUCTION

Phishing attack classically starts by sending an email that appears to come from an enterprise to victims asking them to update or confirm their personal information by visiting a link within the email. Although, phishers are now using several techniques in creating phishing sites, they all use a set of mutual features to create phishing websites since without those features they lose the advantage of deception. This helps us to differentiate between honest and phishy websites based on the features extracted from the visited website.

Overall, two approaches are used in identifying phishing sites. The first one is based on a blacklist (1), in which the requested URL is compared with those in that list. The downside of this approach is that the blacklist usually cannot cover all phishing websites since, within seconds, a new fraudulent website is launched. The second approach is known as heuristic-based methods (2), where several features are collected from the website to categorize it as either phishy or legitimate. In contrast to the blacklist method, a heuristic-based solution can recognize freshly created phishing websites. The accuracy of the heuristic-based methods depends on picking a set of discriminative features that could help in distinguishing the type of website (3). The way in which the features processed also plays an extensive role in classifying websites accurately. Data mining is one of the research fields that can make use of the features extracted from the websites to find patterns as well as relations among them (4). Data mining is important for decision-making since decisions may be made based on the patterns and rules achieved by the data-mining algorithms. Rules are a common representation of data because they are understood easily by humans (5). Normally, the rule takes the form of IF-THEN, for example: IF condition A∧B∧… ∧Z THEN class Ω.

Where "A∧B∧… ∧Z" holds the values of the feature and it is called rule-antecedent, and "Ω" is the predicted class and it is named rule-consequent. The process of detecting unseen knowledge in datasets is represented in terms of rules and is known as rule-induction. Rule-induction eases decision-making because it generates

knowledge that ensures correctness, reliability and completeness (5), as well as reduce the time of knowledge achievement (4). There are two kinds of rules-induction approaches in data-mining : associative approach and classification-rule approach. The use of classification-rules are of concern in this paper. The classification problem goal is to assign each test data to one of the predefined classes in the test dataset. Several studies have been conducted about phishing detection based on website features but these researches were unable to identify precise rules to classify the type of website.

This paper differs from previous researches by proposing a group of features that can be extracted automatically using our own software tool. These features are examined in predicting phishing websites using rules derived from rule-induction algorithms aiming to reduce the false-negative rate i.e. classifying phishing websites as legitimate. Moreover, we showed that extracting features automatically is faster than manual extraction, which in turn increases the dataset size and allows us to conduct more experiments and thus improving the prediction accuracy.

In this article, we try to answer the following research questions:

1. What is the effective minimal set of features that can be utilized in predicting the type of website?
2. How good is rule-based data-mining technique in predicting phishing websites?
3. Which rule based classification technique is more accurate in predicting phishing websites?

This article is structured as follows: Section II defines phishing problem and the damages it causes. Section III discusses related works and highlights different phishing detection methods presented in the literature. Section IV introduces different phishing features, and grouping them into different categories. Finally, in Sections VII, VIII and IX we perform several experiments to assess the significance of the proposed features in detecting phishing website and evaluate different rule based classification algorithms for the same purpose. We conclude in Section XI.

## II. PROBLEM STATEMENT

Phishing websites are fake websites generated by dishonest people to impersonate honest websites. Users may be unable to access their emails or sometimes lose money because of phishing. Predicting and stopping this attack is a critical step toward protecting online trading. The accuracy of predicting the type of the website necessarily depends on the extracted features goodness. Since most users feel safe against phishing attacks if they utilize an anti-phishing tool, this throws a great responsibility on the anti-phishing tools to be accurate in predicting phishing.

In that context, we believe that developing rules of thumb to extract features from websites then utilizing them to predict the type of websites is the key to success in this event.

A report published by "Gartner" (6), which is a research and advisory company shows that phishing attacks continue to escalate. Gartner estimates that theft through phishing attacks costs U.S. banks and credit card issuers an estimated $2.8 billion annually. The Director of Cisco's security-technology-business-unit said (7), "Personalized and targeted attacks that focus on gaining access to more lucrative corporate bank accounts and valuable intellectual property are on the rise".

## III. RELATED WORK

Although quite a lot of of anti-phishing solutions are offered nowadays, most of them are not able to make a high accurate decision thus the false-positive decisions raised intensely.

In this section, we review current anti-phishing methodologies and the features they employ in developing anti-phishing solutions.

One approach employed in (8), is based on experimentally contrasting association classification algorithms, i.e. Clasification Based Association (CBA) ,and Multi-class Classification based on Association classification with other traditional clasification algorithms (C4.5, PART,…etc.). The authors have gathered 27 different

features from various websites and then categorize them into six criteria as shown in Table 1. To evaluate the selected features, the authors conducted experiments using the following data-mining techniques, MCAR (9), CBA (10), C4.5 (11), PRISM (12), PART (4) and JRip (4). The results showed a significant relation between "Domain-Identity" and "URL" features. There was insignificant impact of the "Page Style" on "Social Human Factor" related features on the accuracy.

Table 1 E-BANKING PHISHING CRITERIA

| Category | Phishing Factor Indicator |
|---|---|
| **URL & Domain Identity** | Using IP Address |
| | Request URL |
| | URL of Anchor |
| | DNS Record |
| | Abnormal URL |
| **Security & Encryption** | SSL Certificate |
| | Certification Authority |
| | Abnormal Cookie |
| | Distinguished Names Certificate (DN) |
| **Source Code & Java Script** | Redirect Pages |
| | Straddling Attack |
| | Pharming Attack |
| | Using onMouseOver |
| | Server Form Handler |
| **Page Style & Contents** | Spelling Errors |
| | Copying Website |
| | "Submit" Button |
| | Using Pop-Ups Windows |
| | Disabling Right-Click |
| **Web Address Bar** | Long URL Address |
| | Replacing Similar Characters for URL |
| | Adding Prefix or Suffix |
| | Using the @ Symbol to Confuse |
| | Using Hexadecimal Character Codes |
| **Social Human Factor** | Much Emphasis on Security and Response |
| | Generic Salutation |
| | Buying Time to Access Accounts |

Later in 2010 (13), the authors used the 27 features to build a model based on fuzzy-logic. Although, this is a promising solution it lacks to clarify how the features were extracted from the website, specifically features related to human-factors. Moreover, the rules were established based on human experience, which is one of the problems we aim to resolve in this article. Furthermore, the website was classified into five different

classes i.e. (Very Legitimate, Legitimate, Suspicious, Phishy and Very Phishy), but the authors did not clarify what is the fine line that differentiates between these classes.

Another method proposed in (14), suggested a new way to detect phishing webbehaviorsapturing abnormal behavwebsitesmonstrated by these websites. The authors have selected six structural-features those are: Abnormal URL, Abnormal DNS record, Abnormal Anchors, Server-Form-Handler, Abnormal cookie, and Abnormal SSL-certificate. Once these features and the website identity are known, Support-Vector-Machine classifier "Vapnik's" (15) is used to determine whether the website is phishy or not. The classification accuracy of this method was 84%, which is relatively considered low. However, this method snubs important features that can play key role in determining the legitimacy of a website. This explains the low detection-rate. Nevertheless, this approach does not depend on any previous knowledge of the user or experience in computer security.

A method proposed in (16), suggested utilizing CANTINA (Carnegie Mellon Anti-phishing and Network Analysis Tool) which is content-based technique to detect phishing websites, using the term-frequency-inverse-document-frequency (TF-IDF) information-retrieval measures (17). TF-IDF produces weights that assess the term importance to a document, by counting its frequency. CANTINA works as follow:

1. Calculate the TF-IDF for a given webpage.
2. Take the five highest TF-IDF terms and find the lexical-signature.
3. The lexical-signature is fed into a search engine.

If the N tops searching results having the current webpage, it is considered a legitimate webpage. Otherwise, it is a phishing webpage. N was set to 30 in the experiments. If the search engine returns zero result, thus the website is labelled as phishy, this point was the main disadvantage of using such technique since this would

increase the false-positive rate. To overcome this weakness, the authors combined TF-IDF with some other features; those are Age of Domain, Known-Images, Suspicious-URL, IP-Address, Dots in URL, Forms.

Another approach that utilizes CANTINA with an additional attribute and uses different machine-learning algorithms was proposed in (1). The authors have used 100 phishing websites and 100 legitimate ones in the experiments which is considered limited. The authors have performed three experiments; the first one evaluated a reduced CANTINA features set i.e. (dots in URL, IP-address, suspicious-URL and suspicious-link), and the second experiment involved testing whether the new feature i.e. (domain top-page similarity) is significant enough to play a key role in detecting website type. The third experiment evaluated the results after adding the new suggest feature to the reduced CANTINA features. The result showed that the new feature plays a key role in detecting the website type. The best accurate algorithm was Neural Network with error-rate equals to 7.50%, and Naïve Bayes (NB) gave the worst result with 22.5% error-rate.

In (18), the authors compared a number of learning methods including Support-Vector-Machine, rule-based techniques, decision-trees, and Bayephphishingechniques in detecting phishy emails. A random forest algorithm was implemented in PILFER (Phishing Identification by Learning on Features of Email Received). PILFER detected 96% of the phishing emails correctly with a false-positive rate of 0.1%. Ten email's features displayed are used in the experiments those are IP address URLs, Age of Domain, Non-matching URLs, "Here" Link, HTML emails, Number of Links, Number of Domains, Number of Dots, Containing Javascript, Spam-filter Output.

## IV. CATEGORIZING PHISHING FEATURES

Going back to the approaches presented in Section III, and after reviewing surveys, we were able to derive many features of phishing websites and then categorize them in new groups as shown in Fig. 1.

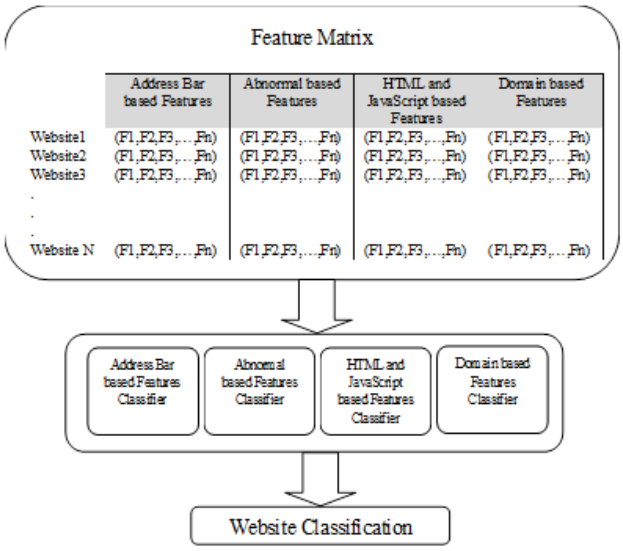


Figure 1 proposed phishing prediction hierarchical model

The features have been categorized based on their impact within a website. Firstly, we examine if a webpage holds text fields, because a phishing webpage asks users to input personal information through those fields. If the webpage has at least one text field we continue to extract other features. Otherwise, the extraction process terminates. To measure the feature significance in detecting phishing we have collected 2500 datasets from the Phishtank (19) and Millersmiles archive (20) using our tool and computed each feature frequency within the dataset in order to reflect the feature importance. In the next section, every feature will be associated with a weight corresponding to the ratio of that feature in the dataset. These frequencies will give us an initial indication of how influential is the feature in a website.

### A. ADDRESS BAR BASED FEATURES

#### *1. IP Address*

If IP address is used as an alternative of a domain name in the URL e.g. 125.98.3.123 or it can be transformed to hexadecimal representation e.g. http://0x58.0xCC.0xCA.0x62, the user can almost be sure someone is trying to steal his personal information. By reviewing our dataset, we find 570 URLs having IP address which

constitutes 22.8% of the dataset. To produce a rule for extracting this feature, we examine the domain part of the URL which lies between "//" and "/", as shown in Fig. 2.

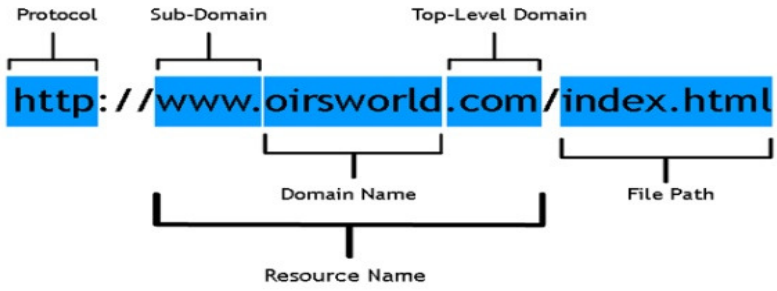


Figure 2 URL Anatomy

Proposed Rule: IF $\begin{cases} \text{IP address exist in URL} \rightarrow \text{Phishy} \\ \text{otherwise} \rightarrow \text{feature} = \text{Legitimate} \end{cases}$

1. *Long URL*

Long URLs commonly used to hide the doubtful part in the address bar. Scientifically, there is no reliable length distinguishes phishing URLs from legitimate ones. As in (21), the proposed length of legitimate URLs is 75. However, the authors did not justify the reason behind their value. To ensure accuracy of our study, we calculated the length of URLs of the legitimate and phishing websites in our dataset and produced an average URL length. The results showed that if the length of the URL is less than or equal 54 characters then the URL classified as "Legitimate". On the other hand, if the URL length is greater than 74 characters then the website is "Phishy". In our dataset, we find 1220 URLs lengths greater than or equal 54 characters, which constitute 48.8%.

Proposed Rule: $\begin{cases} \text{URL length} < 54 \rightarrow \text{Legitimate} \\ \text{URL length} \geq 54 \text{ and } \leq 75 \rightarrow \text{Suspecious} \\ \text{otherwise} \rightarrow \text{Phishy} \end{cases}$

2. *Using @ Symbol*

Using "@" symbol in the URL leads the browser to ignore everything preceding the "@" symbol since the real address often follows the "@" symbol. . After reviewing our dataset, we were able to find 90 URLs having "@" symbol, which constitute only 3.6%.

Proposed Rule: IF $\begin{cases} \text{URL has @ symbol} \rightarrow \text{Phishy} \\ \text{otherwise} \rightarrow \text{Legitimate} \end{cases}$

3. *Prefix or Suffix Separated by "-"*

Dashes are rarely used in legitimate domain-names. Phishers resort to add suffixes or prefixes separated by "-" to the domain names so that users feel they are dealing with a legitimate webpage. 661 URLs having "-" symbol were found in our dataset which constitutes 26.4%.

Proposed Rule: IF $\begin{cases} \text{domain part includes } " - " \text{ symbol} \rightarrow \text{Phishy} \\ \text{otherwise} \rightarrow \text{Legitimate} \end{cases}$

4. *Sub-Domain and Multi Sub-Domains*

Assume that we have the following link http://www.hud.ac.uk/students/portal.com. A domain-name always includes the top-level domain (TLD), which in our example is "uk." The "ac" part is shorthand for academic, ".ac.uk" is called the second-level domain (SLD), and "hud" is the actual name of the domain. We note that the legitimate URL link has two dots in the URL since we can ignore typing "www.". If the number of dots is equal to three then the URL is classified as "Suspicious" since it has one sub-domain. However, if the dots are greater than three it is classified as "Phishy" since it will have multiple sub-domains. Our dataset contains 1109 URLs having three or more dots in domain part, which constitute 44.4%.

Proposed Rule: IF $\begin{cases} \text{dots in the domain part} < 3 \rightarrow \text{Legitimate} \\ \text{else if dots in domain part} = 3 \rightarrow \text{Suspecious} \\ \text{otherwise} \rightarrow \text{feature} = \text{Phishy} \end{cases}$

5. *HTTPS "Hyper Text Transfer Protocol with Secure Sockets Layer" and SSL "Secure Sockets Layer"*

Legitimate websites utilize secure domain-names every time sensitive information is transferred. The existence of https is important in giving the impression of website legitimacy, but it is not enough, since in 2005 more than 450 phishing URLs using https recognized by Netcraft Toolbar Community (22). Therefore, we further checked the certificate assigned with https including the extent of trust of certificate issuer unlike some previous researches, which consider fake https as a valid without checking the certificate of the authority provider. Certificate authorities that are consistently listed among the top names for trust include GeoTrust, GoDaddy, Network Solutions, Thawte, and VeriSign. By reviewing our dataset, we find that the minimum certificate age for the URLs supporting HTTPs protocol was 2 years. In our dataset, we find 2321 URLs does not support https or use a fake https, which constitute 92.8%.

$$\text{Proposed Rule: IF} \begin{cases} \text{use https and trusted issuer and age} \geq 2 \text{ years} \rightarrow \text{Legitimate} \\ \text{using Https and issuer is not trusted} \rightarrow \text{Suspecious} \\ \text{otherwise} \rightarrow \text{Phishy} \end{cases}$$

### B. ABNORMAL BASED FEATURES

1. *Request URL*

For legitimate websites, most of the objects within the webpage are linked to the same domain. For example, if the URL typed in the address bar was http://www.hud.ac.uk/students/portal.com we extract the keyword <src=> from the webpage source code and check whether the domain in the URL is different from that in <src>. If the result is true, the website is classified as "Phishy". To develop a rule for this feature, we calculated the ratio of URLs in source code that have different domain than the domain typed in the address bar. By reviewing our dataset, we find that the legitimate websites have in the worst case 22% of its objects loaded from different domains, whereas for phishing websites the ratio was in best case was 61%. Thus, we assumed that if the ratio is less than 22% then the website is considered "Legitimate" else if the ratio is

between 22% and 61% then the website considered “Suspicious”. Otherwise, the website is considered “Phishy”. In this feature, we computed the feature existence rate not the number of feature existence; since the number of request URL in the website varies. The dataset contains 2500 URLs having this feature, which constitute 100 %.

$$\text{Proposed Rule: } \begin{cases} \text{request URL \%} < 22\% \rightarrow \text{Legitimate} \\ \text{request URL\%} \geq 22\% \text{ and} < 61\% \rightarrow \text{Suspecious} \\ \text{otherwise} \rightarrow \text{Phishy} \end{cases}$$

2. *URL of Anchor*

An anchor is an element defined by the <a> tag. This feature is treated exactly as “Request URL”. By reviewing our dataset, we find that the legitimate websites have in the worst case 31% of its anchor-tag connected to a different domain, whereas for phishing websites we find that the ratio was 67% in best case. Thus, we assumed that if the ratio is less than 31% then the website is considered “Legitimate” else if the ratio is between 31% and 67% then the website considered “Suspicious”. Otherwise, the website is considered “Phishy”. By reviewing our dataset, we find 581 URLs having this feature, which constitute 23.2%.

$$\text{Proposed Rule: } \begin{cases} \text{URL of anchor \%} < 31\% \rightarrow \text{Legitimate} \\ \text{URL of anchor \%} \geq 31\% \text{ and} \leq 67\% \rightarrow \text{Suspecious} \\ \text{otherwise} \rightarrow \text{Phishy} \end{cases}$$

3. *Server Form Handler (SFH)*

SFH that contains empty string or “about:blank” are considered doubtful since an action should be taken upon submitted information. In addition, if the domain-name in SFH-s is different from the domain of the webpage this gives an indication that the webpage is suspicious because the submitted information is rarely handled by external domains. In our dataset, we find 101 URLs having SFHs, which constitutes only 4.0%.

$$\text{Proposed Rule: } \begin{cases} \text{SFH is "about: blank" or an empty} \rightarrow \text{Phishy} \\ \text{SFH refers to a different domain} \rightarrow \text{Suspecious} \\ \text{otherwise} \rightarrow \text{Legitimate} \end{cases}$$

*4. Abnormal URL*

This feature can be extracted from WHOIS database (19). For a legitimate website, identity is typically part of its URL. 412 URLs having this feature were founded in our dataset, which constitutes 16.4%.

Proposed Rule: IF $\begin{cases} \text{the host name is not included in URL} \rightarrow \text{Phishy} \\ \text{otherwise} \rightarrow \text{Legitimate} \end{cases}$

### C. HTML AND JAVASCRIPT BASED FEATURES

*1. Redirect Page*

Open redirects found on websites are liable to be exploited by phishers to create a link to their site. In our dataset, we find that the maximum number of redirect pages in the phishing websites was three, whereas this feature is rarely used in legitimate websites since we find only 21 legitimate website having this feature and it is used for one time only in those websites. Thus if the redirection number is less than 2 then we will assign “Legitimate”, else if the redirection number is greater than or equal 2 and less than 4 then we will assign "Suspecious", otherwise we will assign “Phishy”. 249 URLs having redirect-page were encountered in our phishing dataset, which constitute 10%.

Proposed Rule: $\begin{cases} \text{redirect page \#} \leq 1 \rightarrow \text{Legitimate} \\ \text{redirect page \#} > 1 \; and \; < 4 \rightarrow \text{Suspecious} \\ \text{otherwise} \rightarrow \text{Phishy} \end{cases}$

*2. Using onMouseOver to Hide the Link*

Phishers may use JavaScript to display a fake URL in the status bar to the users. To extract this feature we must explore the webpage source code particularly the “onmouseover” event and check if it make any changes on the status bar. 496 URLs having this feature were founded in our dataset, which constitutes 20%.

Proposed Rule: $\begin{cases} \text{onmouseover change the status bar} \rightarrow \text{Phishy} \\ \text{it does not change status bar} \rightarrow \text{Suspecious} \\ \text{otherwise} \rightarrow \text{Legitimate} \end{cases}$

*3. Disabling Right Click*

Phishers use JavaScript to disable the right-click function, so that users cannot view and save the source code. This feature is treated exactly as "Using onMouseOver to hide the Link". However, for this feature, we will search for event "event. button==2" in the source code and check if right click is disabled. We find this feature 40% times in our dataset, which constitutes 1.6%.

Proposed Rule: $\begin{cases} \text{right click disabled} \rightarrow \text{Phishy} \\ \text{right click showing an alert} \rightarrow \text{Suspecious} \\ \text{otherwise} \rightarrow \text{Legitimate} \end{cases}$

*4. Using Pop-up Window*

It is unusual to find a legitimate website asking users to submit their credentials through a popup window. 227 URLs were founded in our dataset in which users credential submitted through a popup window, which constitutes 9.1%.

Proposed Rule: $\begin{cases} \text{Not using popup} \rightarrow \text{Legitmate} \\ \text{otherwise} \rightarrow \text{feature} = \text{Phishy} \end{cases}$

### D. DOMAIN BASED FEATURES

*1. Age of Domain*

This feature can be extracted from WHOIS database (23). In our dataset, we find that some domains host several phishy URL in several time slots. The blacklist may succeed in protecting the users if it works on the domain level not on the URL level i.e. add the domain-name to the blacklist not the URL address. However, (24) find that 78% of phishing domains were in fact hacked domains, which already serve a legitimate website. Thus, blacklisting those domains will in-turn adds the legitimate websites to blacklist as well. Even though the phishing website has moved from the domain, legitimate websites may be left on blacklists for a long time; causing the reputation of the legitimate website or organization to be harmed. Some blacklists such

as “Google’s Blacklist” need on average seven hours to be updated (25). By reviewing our dataset, we find that the minimum age of the legitimate domain was 6 months. For this feature, if the domain created less than six months, it is classified as “Phishy”; otherwise, the website is considered “Legitimate”. In our dataset, 2392 URLs created less than 6 months, which constitute 95.6%.

Proposed Rule: $\begin{cases} \text{age of domain is} \geq 6 \text{ months} \rightarrow \text{Legitimate} \\ \text{otherwise} \rightarrow \text{Phishy} \end{cases}$

2. *DNS Record*

For phishing sites, either the claimed identity in not recognized by the WHOIS database (19) orfoundedcord of the hostname is not founded (14). If the DNS record is empty or not found then the website is classified as “Phishy”, otherwise it is classified as “Legitimate”. 160 URLs were found in our dataset where the DNS record is not found, and that constitute 6.4%.

Proposed Rule: $\begin{cases} \text{no DNS record for the domain} \rightarrow \text{Phishing} \\ \text{otherwise} \rightarrow \text{Legitimate} \end{cases}$

3. *Website Traffic*

This feature measures the popularity of the website by determining the number of visitors and the number of pages they visit. However, since phishing websites lives for a short period-of-time thus they may not be recognized by the Alexa database (26). By reviewing our dataset, we find that in worst-case legitimate websites ranked among the top 150,000. Therefore, if the domain has no traffic or not being recognized by the Alexa database it is classified as “Phishy” otherwise if the website ranked among the top 150,000 it is classified as “Legitimate” else it is classified as “Suspicious”. This feature constitutes 89.2% of the dataset since it appears 2231 times.

Proposed Rule: $\begin{cases} \text{webpage rank} < 150{,}000 \rightarrow \text{Legitimate} \\ \text{The Ranking} > 150{,}000 \rightarrow \text{Suspecious} \\ \text{otherwise} \rightarrow \text{Phish} \end{cases}$

## V. PREPARING FOR EXPERIMENTS

To conduct experiments of producing new rules related to phishing, some preparatory steps must be taken as follows:

- Dataset preparation: A set of phishing websites was collected from Phishtank (19), which is a free community site where users can submit, verify, track and share phishing data. In addition, we utilized the Millersmiles (20), which is considered a prime source of information about spoof emails and phishing scams. The legitimate websites were collected from yahoo directory (27) and starting point directory (28). We collected 2500 phishing URLs, and 450 legitimate ones.
- Address bar features: A JavaScript program was built to extract all features related to the address bar.
- Abnormal based features: A simple PHP script was developed to extract those features since these features deal with servers and require a connection to external domains such as the WHOIS database (23).
- HTML and JavaScript based features: A JavaScript program was built to extract these features.
- Domain based features: These features can be extracted from the WHOIS database (23), and from Alexa.com (26). Furthermore, we developed a PHP script to extract these features, as shown in Fig. 3.

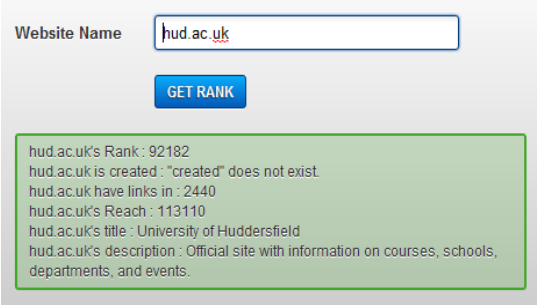


Figure 3 Result of Alexa Query

## VI. FEATURES POPULARITY IN DESIGNING PHISHING WEBSITES

To measure which features are significant in designing phishing websites. We calculated the frequencies for each feature in our datasets, as shown in Table 2. The results showed that the "Request URL" feature is the most popular one since it exists in all 2500 data elements, followed by "Age of Domain" which presented in 2390 data elements. The next popular feature is "HTTPS and SSL". The lowest popular feature is "Disabling Right Click" feature, which appeared only forty times, followed by "URL having @ symbol" which constituted 3.6% of the dataset.

Table 2 Feature popularity in designing phishing websites

| Feature | Frequency | Percentage |
|---|---|---|
| **Using the IP Address** | 570 | 22.8% |
| **Long URL** | 1220 | 48.8% |
| **URL's having @ Symbol** | 90 | 3.6% |
| **Adding prefix or Suffix Separated by (-) to the domain** | 661 | 26.4% |
| **Sub Domain and Multi Sub Domain** | 1109 | 44.4% |
| **HTTPS and SSL** | 2321 | 92.8% |
| **Request URL** | 2500 | 100% |
| **URL of Anchor** | 581 | 23.2% |
| **Server Form Handler** | 101 | 4.0% |
| **Abnormal URL** | 412 | 16.4% |
| **Redirect Page** | 249 | 10.0% |
| **Using onMouseOver** | 496 | 20.0% |
| **Disabling Right-Click** | 40 | 1.6% |
| **Using pop-up Window** | 227 | 9.1% |
| **Age of Domain** | 2392 | 95.6% |
| **DNS Record** | 160 | 6.4% |
| **Website Traffic** | 2231 | 89.2% |

## VII. THE COMPARED RULE BASED CLASSIFICATION ALGORITHMS

In this section, we compare different rule-based classification algorithms, each of which utilizes a different methodology in producing knowledge. The first algorithm is C4.5 (11), which extracts decision-tree from a dataset based on information theory. C4.5 utilizes a divide-and-conquer methodology to develop decision-trees. The second algorithm is RIPPER (4), which adopts a separate-and-conquer technique. The third algorithm is PRISM (12), which is classified under the covering algorithms family. Finally, we utilize

Classification Based on Association algorithm (CBA) (10), wthe Apriori algorithmentation of Apriori algorithm (29). CBA is based on finding the frequent items by passing over the dataset many times aiming to find set of items with support greater than the minimum support threshold. Then after finding the frequent items it produces rules passes the minimum confidence for each frequent item. The support of a rule indicates how frequently the items in the rule's body inside the training dataset. The confidence of a rule represents its strength and is defined the probability of both the rule's antecedent and the consequent together in the training dataset divided by the frequency of the rule's antecedent.

## VIII. EXPERIMENTS

We compare the algorithm's performance for each feature category in our model Fig. 1. We conduct the experiments using the WEKA tool (30), which is an open source data-mining application created in Java at the Waikato University (4). As we mentioned earlier we have 450 legitimate websites, further we randomly picked 450 URLs from phishing dataset thus we have 450-phishing website and 450-legitimate website in our training dataset. Fig. 4 summarizes the prediction error-rate produced by considered algorithms for each dataset.

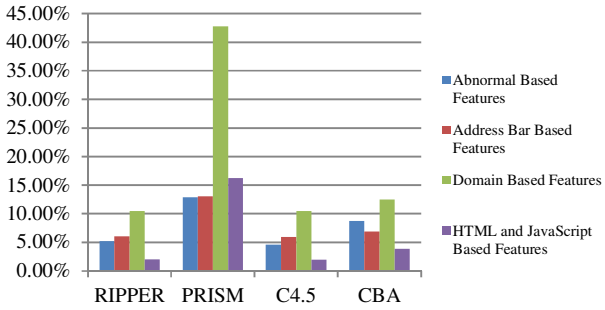


Figure 4 Prediction error rate summary

The results showed that C4.5 algorithm outperforms RIPPER, PRISM and CBA in predicting the class for "Abnormal based dataset", "Address Bar based dataset" and "HTML and JavaScript based dataset" where for

"Domain based dataset" C4.5 and RIPPER have the same error-rate. However, by computing the average error-rate for each algorithm, we noticed that C4.5 outperform all algorithms with 5.76% average error-rate, followed by RIPPER with 5.94% average error-rate whereas the highest average error-rate was achieved by PRISM with 21.24%. Overall, the prediction accuracy obtained from all algorithms considered acceptable and that reflects the goodness of our features in predicting the website class.

## IX. REDUCED FEATURES EXPERIMENTS

The work in this section aims to reduce the number of features in order to reduce the runtime of the data-mining algorithm, as well as nominating the least number of features that can be used to predict phishing websites. In addition, selecting features may eliminate the noise in features, which occurs whenever there are irrelevant features presented within the training dataset, which in turn causes an increase in the classification errors (17). A frequently used metric to evaluate features for relevance for the classification task is Chi-Square (17). Fortunately, WEKA facilitates this method. After evaluating the features using Chi-Square, we find that the best features that may be used to predict phishing websites are:

"*Request URL, Age of Domain, HTTPS and SSL, Website Traffic, Long URL, Sub Domain and Multi Sub Domain, Adding prefix or Suffix Separated by (-) to Domain, URL of Anchor and Using the IP Address*".

We assess the prediction accuracy by means of the same classification algorithms used in section VII. From the results, we noticed that even PRISM has a good prediction-rate, knowing that no rule pruning had taken place while producing the rules, which reflects how good these features are in classifying the websites.

However, digging deeply into the rules produced by each algorithm, we noticed that all algorithms generated the following rule:

*If HTTPS and SSL = High*
*And Age of Domain =High*
*Then Phishing*

This reflects the importance of “HTTPS and SSL” in predicting phishing websites, as well as “Age of Domain”. Going back to the rules proposed to extract “HTTPS and SSL”, the modification on how to extract this feature was very effective.

## X. CONCLUSION:

This article investigated the features that are effective in detecting phishing websites. These features extracted automatically without any intervention from the users and using computerized developed tools. We managed to collect and analyze 17 different features that distinguish phishing websites from legitimate ones. Further, we developed a new rule for each feature. These rules can be useful in applications related to discovering phishing websites based on their features. After performing frequency analysis for each feature, the results showed that “Request URL” is the most popular feature in creating phishing websites since it appears in all dataset cases, followed by “Age of Domain”, which was presented in 2392 dataset cased. The next popular feature is “HTTPS and SSL” with a frequency-rate of 91%. Several experiments have been conducted using different rule-based classification algorithms to extract new hidden knowledge that can help in detecting phishing websites. The experiments showed that C4.5 algorithm outperformed RIPPER, PRISM and CBA in term of accuracy. Furthermore, an experiment conducted after selecting the most effective features in predicting phishing websites. The results showed that we could improve the prediction accuracy relying only on nine features, those are: “Request URL, Age of Domain, HTTPS and SSL, Website Traffic, Long URL, Sub Domain and Multi Sub Domain, Adding prefix or Suffix Separated by (-) to Domain, URL of Anchor and Using the IP Address”. After conducting the experiments on the nine chosen features, the error-rate has decreased for all algorithms. Precisely, CBA algorithm has the lowest error-rate with 4.75%.

In the near future, we will use the rules produced by different algorithms to build a tool that is integrated with a web browser to detect phishing websites on real time and warn the user of any possible attack.

## REFERENCES


1. Nuttapong Sanglerdsinlapachai , Arnon Rungsawang. Using Domain Top-page Similarity Feature in Machine Learning-based Web. In Third International Conference on Knowledge Discovery and Data Mining; 2010: IEEE. p. 187-190.

2. Sophie GP, Gustavo GG, Maryline L. Decisive Heuristics to Differentiate Legitimate from Phishing Sites. In 2011 Conference on Network and Information Systems Security; 2011: IEEE. p. 1-9.

3. Guang X, Jason o, Carolyn P R, Lorrie C. CANTINA+: A Feature-rich Machine Learning Framework for Detecting Phishing Web Sites. ACM Transactions on Information and System Security. 2011 Sep: p. 1-28.

4. Witten IH, Frank E. Data mining: practical machine learning tools and techniques with Java implementations. New York, NY, USA:; March 2002.

5. H DJ. Rule induction-machine learning techniques. Computing & Control Engineering Journal. 1994 October: p. 249-255.

6. Gartner, Inc. [Online]. Available from: http://www.gartner.com/technology/home.jsp.

7. Lennon, M. Security Week. [Online].; 2011. Available from: http://www.securityweek.com/cisco-targeted-attacks-cost-organizations-129-billion-annually.

8. Aburrous, M , Hossain, M. A. , Dahal, K. , Fadi, T. Predicting Phishing Websites using Classification Mining Techniques. In Seventh International Conference on Information Technology.; 2010; Las Vegas, Nevada, USA.: IEEE. p. 176-181.

9. Thabtah F, Peter C, Peng Y. MCAR: Multi-class Classification based on Association Rule. In The 3rd ACS/IEEE International Conference on Computer Systems and Applications; 2005. p. 33.

10. Hu K, Lu Y, Zhou L, Shi C. Integrating Classification and association rule Mining. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98, Plenary Presentation); 1998; New York, USA: Springer-Verlag. p. 443 - 447.

11. Quinlan JR. Improved use of continuous attributes in c4.5. Journal of Artificial Intelligence Research. 1996;: p. 77-90.

12. Cendrowska J. PRISM: An algorithm for inducing modular rule. International Journal of Man-Machine Studies. 1987;: p. 349-370.

13. Aburrous , Hossain MA, Dahal K, Thabtah F. Intelligent phishing detection system for e-banking using fuzzy data mining. Expert Systems with Applications: An International Journal. 2010 December: p. 7913-7921.

14. Pan , Ding. Anomaly Based Web Phishing Page Detection. In In ACSAC '06: Proceedings of the 22nd Annual Computer Security Applications Conference.; Dec. 2006: IEEE. p. 381-392.

15. Cortes C, Vapnik V. Support-Vector Networks. Machine Learning. Sept.1995;: p. 273 - 297.

16. Zhang , Hong , Cranor. CANTINA: A Content-Based Approach to Detect Phishing Web Sites. In Proceedings of the 16th World Wide Web Conference; May, 2007.

17. Manning C, Raghavan , Schütze H. Introduction to Information Retrieval: Cambridge University Press; 2008.

18. Sadeh N, Tomasic A, Fette I. Learning to detect phishing emails. Proceedings of the 16th international conference on World Wide Web. 2007: p. 649-656.

19. PhishTank. [Online].; 2006 [cited 2011 November 25. Available from: http://www.phishtank.com/.

20. millersmiles. millersmiles. [Online].; 2011 [cited 2011 October. Available from: http://www.millersmiles.co.uk/.

21. Horng SJ, Fan P, Khan MK, Run RS, Lai JL, Chen RJ, et al. An efficient phishing webpage detector. Expert Systems with Applications: An International Journal. 2011; 38 (10): p. 12018-12027.

22. More than 450 Phishing Attacks Used SSL in 2005. [Online]. [Cited 2012 March 8. Available from: http://news.netcraft.com/archives/2005/12/28/more_than_450_phishing_attacks_used_ssl_in_2005.html.

23. WhoIS. [Online]. Available from: http://who.is/.

24. Rasmussen R, Aaron G. Global Phishing Survey: Trends and Domain Name Use 2H2009 [Survey]. Lexington; 2010. Available from: http://anti-phishing.org/reports/APWG_GlobalPhishingSurvey_2H2009.pdf.

25. Ask Sucuri. Security Blog. [Online].; 2011. Available from: http://blog.sucuri.net/2011/12/ask-sucuri-how-long-it-takes-for-a-site-to-be-removed-from-googles-blacklist-updated.html.

26. Alexa the Web Information Company. [Online]. [Cited 2012 January 26. Available from: http://www.alexa.com/.

27. Yahoo Directory. [Online]. Available from: http://dir.yahoo.com/.

28. Starting Point Directory. [Online]. Available from: http://www.stpt.com/directory/.

29. Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules. VLDB '94 Proceedings of the 20th International Conference on Very Large Data Bases. 1994;: p. 487-499.

30. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. Waikato Environment for Knowledge Analysis. [Online]. Available from: http://www.cs.waikato.ac.nz/ml/weka/.