International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 8 ǁ August 2022 ǁ PP. 602-611

Using Machine Learning Approaches to Verify the Security of Short URLs

WC Yeh [1], Zhiyuan Chen [2], Zongping Wei [3], Hsiuchi Dai [4], Akshay Kumar [5]
[1] *Department of Electrical Engineering, National Tsing Hua University, Taiwan*
Email: yeh@iclabnthu.org
[2] *Department of Mathematics, University of California, Berkeley, USA*
Email: zhiyuanchen2000@berkeley.edu
[3] *Department of Electrical Engineering, Wuhan University of Technology, China*
Email: zongpingwei@gmail.com
[4] *Department of Electrical Engineering, Wuhan University of Technology, China*
Email: hcdai881226@hotmail.com
[5] *Department of Electronic and Electrical Engineering, Indian Institute of Technology Bombay, India*
Email:akshaykumar@gmail.com
Corresponding Author:WC Yeh

Abstract
On the Internet, the website address is the standard resource address. However, it is relatively difficult to process long URLs. When the URL is composed of special characters, it is easily seen as malicious and users would not be able to use it. Therefore, URL shortening services have become popular. It can shorten URLs so that they become easily readable and is under the length limit. Whereas, in recent times, people have been exploiting URL shortening services to package phishing websites. By the redirection functionality of short URLs, no anti-virus software can detect it and this makes it difficult for users to identify them, leading users' private information to be leaked. The model proposed in this article restores the short URL into the original URL and makes comparison with the online phishing website database by the extracted values of the original URL. Then, machine learning techniques are employed to make predictions, allowing users to identify the short URL with trust.

Keywords: *Short URL, Phishing, Machine Learning, Weka.*

Date of Submission: 18-08-2022 Date of acceptance: 02-09-2022

I. INTRODUCTION

The URL shortening service was patented in 2000 [8], and has become a favorite tool for Internet users. Through the URL shortening service, users can shorten the URL they want to share into a fixed-length URL. In this way, the layout of the transmitted message can be more concise, and can also effectively meet the length restriction of each post set forth by social media. For example, Twitter limits each tweet to only 140 characters. The URL of the file stored in the cloud service is a very long string of text. Using the URL shortening service can effectively shorten the URL to allow users to receive messages more effectively. However, owing to the fact that a shortened URL has a fixed length, it is easy to arrange and combine it by brute force attack, allowing others to tamper with the file content in the cloud space, import malicious programs, and even steal the user's identity information [6].

According to the report of Trend Micro Global Technical Support and R&D Center, short URLs were found to be used to send spam in 2010. In 2012, a Skype worm was found to spread rapidly using short URLs by pointing users to a malicious file, and users who clicked the link become the puppet of the Botnet. In 2014, an ACH scam email emerged, and when users clicked on the link, they were directed to a .zip file containing a malicious program that poisoned the user's device [11]. Today, short URLs are often used to package phishing websites. Phishing websites use the redirection feature of short URLs to easily evade detection by antivirus software.

Indeed, although short URLs bring convenient services, malicious use of it also brings negative effects to users. Therefore, this article proposes a model for verifying the security of short URLs. First, the short URL is restored. The original URL is compared through three online database platforms, namely PhishTank, OpenPhish, and PhishRepo that search for phishing websites. Second, multiple characteristics are extracted from the original URL and compared with the phishing website data recorded in the UCI Machine Learning Repository. Then, the Weka software will be executed to generate three different machine learning prediction

models for comparison, providing reports for users to decide whether the website is a phishing website and determine the security of the website.

The structure of this article is as follows. In second 2, we will discuss and analyze the existing literature; then the experimental method of this research model will be described in section 3; in section 4, we will present the experimental results and analysis; Finally, some conclusions and future possible researches are proposed in section 5.

II. LITERATURE REVIEW

URL shortening is a technique that maps an original URL to a short URL that could redirect users back to the original URL. The concept of short URLs was originally designed to prevent users from destroying the integrity of the original complex URLs in the process of copying URL characters, and by then there was no symbol that could wrap long URLs [9].

There are many URL shortening services on the Internet today. The service will generate a fixed-length URL from the original URL entered by the user, and the URL is composed of three parts, such as: "https://tinyurl.com/65cftsef6ej " (this URL is produced by the original URL of Google https://www.google.com through the TINYURL shortening service). It is composed of three parts: https, tinyurl.com and nilusef6ej: https stands for the protocol, tinyurl.com is the domain name of the service, and 65cftsef6ej is a fixed-length hash code. The generation of the short URL can succinct the message, making it easier for users to read other description text, and not to occupy too much information space due to the length of the original URL.

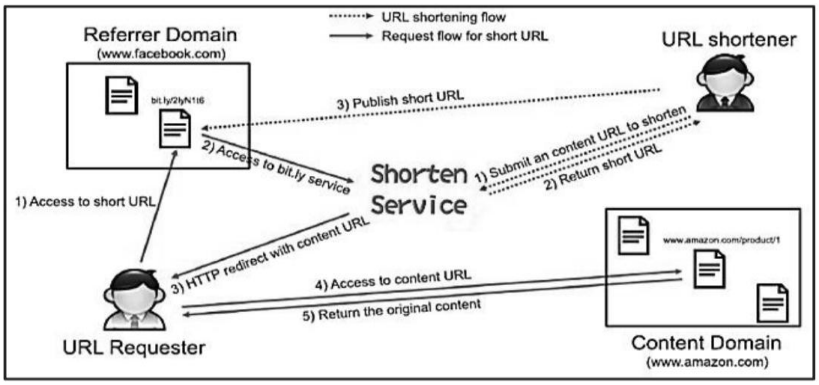


Figure 1: The shortening process of URL

The shortening process of URL can be divided into four parts: the user who shortens the URL (URL shortener), the user who clicks the URL shortener(URL Requester), the Referrer Domain that provides the URL shortening service, and the Content Domain of the original URL, as shown in Figure 1[5]. The URL shortener first transmits the URL to be shortened to the URL shortening service platform, and the URL shortening service platform will return a short URL to the user, and the user can share and transmit the short URL to others. When another URL Requester receives the short URL, the short URL will retrieve the original URL from the service platform it belongs to, and then redirect the user to the original URL for follow-up service.

In the past, many scholars have conducted research on short URLs:

1. Demetris analyzed the use of short URLs on the Twitter social media platform in 2011, including daily statistics, the number of reposts (tweets), and the number of clicks.
2. Klien and Strhmaier studied the malicious abuse of short URLs in 2012. Since short URLs can hide the content of the original URL, the scholar studied the abuse of short URLs used to send spam.
3. In 2011, Chhabra studied how phishing websites share through the Bit.ly URL shortening service and compared the data set collected by PhishTank (phishing website database).
4. Wang proposed a Twitter-based social platform in 2013 to detect whether there are spam emails with short URLs in emails.
5. In 2013, Maggi collected short URLs that were used within two years, and found that short URLs were used maliciously. The damage caused by the threat was actually not as mature as the threat caused by long URLs.
6. Nikiforakis reported in 2014 that short URLs based on advertisements are easily abused by malicious software to confuse users to obtain sensitive information.

7. Scholars such as Daejin Choi collected 180 million short URLs on Bit.ly's URL shortening platform in 2018, analyzed the information of 4.2 billion clicks, and found that the original URL content of most of the generated short URLs is invalid.

Based on the above literature, it can be seen that the current research by scholars has not proposed an effective mechanism to identify whether a short URL is safe. Therefore, this paper proposes a security mechanism that firstly compares the short URL with the phishing URL in the public database, then extracts the special value of the URL, and finally uses the model trained by machine learning techniques to predict whether the short URL is a malicious phishing URL.

III. THE PROPOSED METHOD

The model proposed in this study is to be divided into three steps (as shown in Figure 2). The first step is to restore the short URL into the original URL; the second step is to compare the original URL with online public websites such as PhishTank, OpenPhish, and PhishRepo to determine whether it has been reported as a phishing URL; the third step is to extract 30 special characteristics of the URL and use the three prediction models generated by Weka using machine learning techniques, namely Naive Bayes, J48, and Random forest, to make predictions on the URLs collected, and then compare the prediction accuracy of the three machine learning methods.

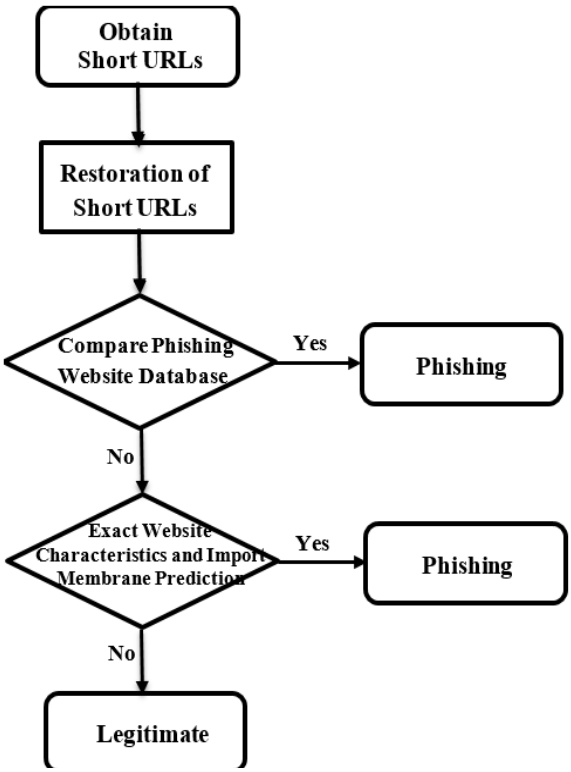


Figure 2: The experimental process

3.1. Characteristics of phishing websites

According to the definition of APWG (Anti-Phishing Working Group), a phishing website is an illegal website that uses social engineering technology to steal consumers' personal identity data and financial account credentials. Social engineering, on the other hand, refers to tricking the user into believing that they are connecting to a trusted website, such as using fake emails and deceptive messages to direct consumers to fake websites. Recipients, therefore, disclose their own financial information, such as usernames and passwords [4]. We use and extract 30 special characteristics from the website data collected in 2015 from the UCI Machine Learning Repository (a well-known foreign public database), which contains 11,055 valid website data sets, of which 6,157 are phishing websites and 4,898 websites are legitimate (non-phishing websites) to judge whether they are legitimate websites or phishing websites, as listed in Table 1 [10].

Table 1: 30 Characteristics of phishing websites.

Website	Field Description
A	Using the IP Address
B	Long URL to Hide the Suspicious Part
C	Using URL Shortening Service “TinyURL”
D	URL’s having “@” Symbol.
E	Redirecting using “//”.
F	Adding Prefix or Suffix Separated by (-) to the Domain.
G	Sub Domain and Multi Sub Domains
H	HTTPS
I	Domain Registration Length.
J	Favicon
K	Using Non-Standard Port
L	The existence of “HTTPS” Token in the Domain Part of the URL.
M	Request URL.
N	URL of Anchor
O	Links in <Meta>, <Script>, and <Link> tags.
P	Server Form Handler (SFH).
Q	Submitting Information to Email.
R	Abnormal URL.
S	Website Forwarding
T	Status Bar Customization
U	Disabling Right Click.
V	Using Pop-up Window
W	Iframe Redirection.
X	Age of Domain
Y	DNS Record
Z	Website Traffic
AA	PageRank
AB	Google Index
AC	Number of Links Pointing to Page
AD	Statistical-Reports Based Feature

The thirty characteristics of websites are categorized into four categories as "URL-based", "domain-based", "domain/syntax-based", and "abnormal", as follows:

(1) URL-based:
A. Using the IP Address
B. Long URL to Hide the Suspicious Part
C. Using URL Shortening Service “TinyURL”
D. URL having “@” Symbol
E. Redirecting using “//”
F. Adding Prefix or Suffix Separated by (-) to the Domain.
G. Sub Domain and Multi Sub Domains
H. HTTPS
I. Domain Registration Length
J. Favicon
K. Using Non-Standard Port
L. The existence of “HTTPS” Token in the Domain Part of the URL.

(2) Domain-based:
M. The content of the page is loaded by another domain
N. The syntax of <a> in website content exceeds the proportion
O. A large number of <Meta>, <Script>, and <Link> in the source code
P. The server's form handler contains a blank string
Q. The form allows users to submit personal information
R. Use Whois for registration check

(3) Domain/syntax-based:
S. The number of times a site redirects to another domain
T. Whether the website is embedded with fake Javascript
U. Whether the website disables mouse clicks
V. Whether there is a pop-up window on the website
W. Is the Iframe in the website hidden?
X. Lifetime of domain existence
Y. Whether the domain is exist in Whois

(4) Abnormal:
Z. Whether the site's traffic is identified by the Alexa database
AA. Ranking of web pages
AB. The importance of web pages on the Internet
AC. Whether the website is searchable on Google
AD. Whether the domain and IP have been notified of suspicious websites

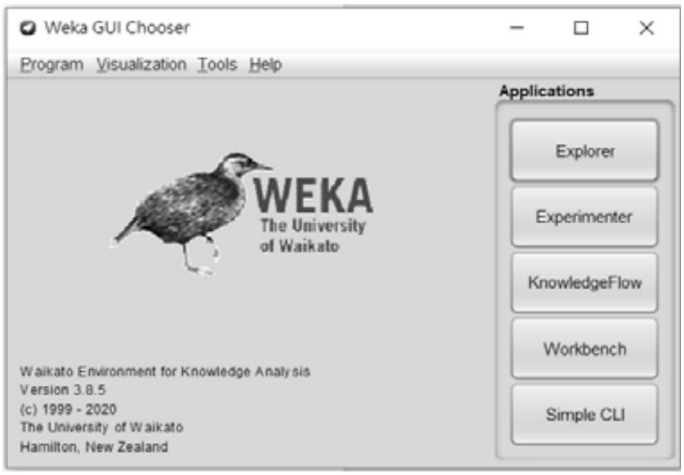


Figure 3: Weka opening screen

3.2. Weka

Weka was developed by the University of Waikato in New Zealand and consists of the initials of the phrase Waikato Environment for Knowledge Analysis. Weka is a set of software that provides data mining and machine learning tools, including data preprocessing tools, classification tools, regression analysis, etc., and it can also visualize data. This software is compiled with Java programming language [7].The software interface includes data mining methods: preprocessing, classification, grouping, association rules, and selection of data attributes (as shown in Figure 4). Weka can perform data analysis on the data set in ".arff" file format, find out the correlation between the data, use machine learning models to predict new data, or use different machine learning methods to make predictions with more suitable models.

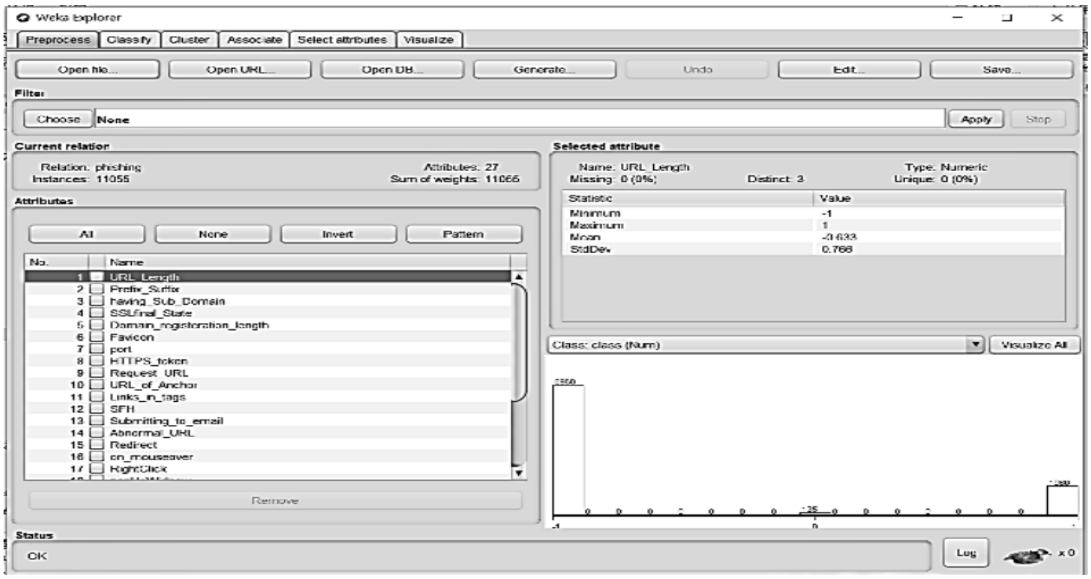


Figure 4: Weka noise interface

In this article, the 30 special characteristics extracted from each of the aforementioned websites are first normalized: they are denoted by the values of "1", "0" and "-1", which respectively represent the “Phishing website", “Suspected Phishing Website”, and “Legitimate Website”. Then, through Weka, the phishing website data set is used to build a prediction model with three kinds of machine learning algorithms, which are described as follows:

(1) **Naïve Bayes**:

Naive Bayes classification algorithm is a classification method based on Bayesian theorem, which is derived from conditional probability. The conditional probability refers to "the probability that event B occurs given condition A", which is equal to the probability of both A and B occurs divided by the probability that A occurs, as in formula (1):

$$P(B \mid A) = \frac{P(A \text{I } B)}{P(A)} \qquad (1)$$

Using Naïve Bayes to build a model of the phishing website data set, the accuracy rate is 94.0554%, as shown in Figure 5.

```
=== Summary ===

Correctly Classified Instances        2310               94.0554 %
Incorrectly Classified Instances       146                5.9446 %
Kappa statistic                          0.8797
Mean absolute error                      0.0766
Root mean squared error                  0.2137
Relative absolute error                 15.5145 %
Root relative squared error             42.9911 %
Total Number of Instances             2456

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.932     0.053     0.934      0.932     0.933     0.880    0.987     0.984     1
               0.947     0.068     0.946      0.947     0.946     0.880    0.987     0.989     -1
Weighted Avg.  0.941     0.061     0.941      0.941     0.941     0.880    0.987     0.987

=== Confusion Matrix ===

   a    b   <-- classified as
1020   74 |    a = 1
  72 1290 |    b = -1
```

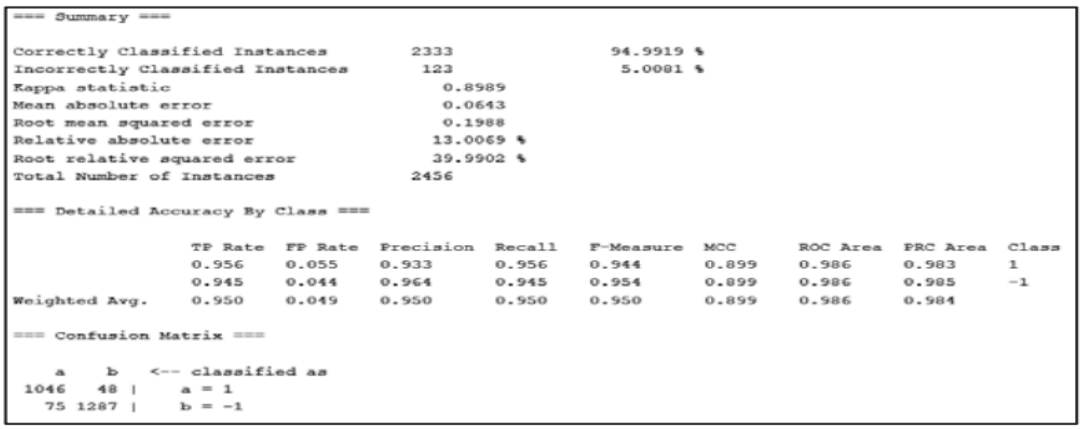
Figure 5: Modeling the characteristics of the phishing website database with the Naive Bayes algorithm

(2) **J48**:

J48 is a decision tree algorithm, also known as C4.5 algorithm. It is a divide-and-conquer strategy based on a top-to-bottom recursion. First, an attribute is selected and placed at the root node, a branch is then generated,

dividing the instances into multiple subsets, and each subset corresponds to a branch of the root node. This process is recursively repeated on each branch until all instances have the same classification. The J48 core algorithm inherits the advantages of the ID3 algorithm, while improving the following aspects:

- Use the information gain rate to select attributes, which overcomes the tendency to select attributes with more values when selecting attributes with information gain.
- Pruning is possible during tree construction.
- Able to discretize continuous attributes.
- Ability to process incomplete data.



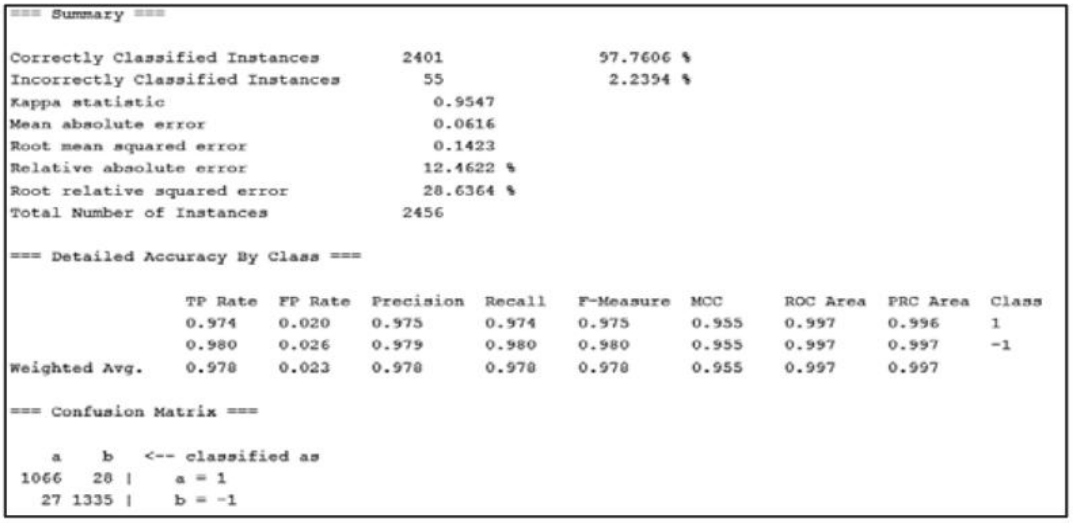
```
=== Summary ===

Correctly Classified Instances        2333               94.9919 %
Incorrectly Classified Instances       123                5.0081 %
Kappa statistic                          0.8989
Mean absolute error                      0.0643
Root mean squared error                  0.1988
Relative absolute error                 13.0069 %
Root relative squared error             39.9902 %
Total Number of Instances             2456

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.956     0.055     0.933     0.956     0.944     0.899     0.986     0.983     1
                 0.945     0.044     0.964     0.945     0.954     0.899     0.986     0.985     -1
Weighted Avg.    0.950     0.049     0.950     0.950     0.950     0.899     0.986     0.984

=== Confusion Matrix ===

    a    b   <-- classified as
 1046   48 |    a = 1
   75 1287 |    b = -1
```

Figure 6: Modeling the characteristics of the phishing website database with the J48 algorithm

(3)**Random Forest**:

Random Forest is the combination of multiple CART trees (Classification and Regression Tree), alongside randomly allocated training data that greatly improve the final calculation result. The theory of the random forest algorithm is that according to the law of large numbers, a forest is composed of k decision trees, and k random vectors are also generated, and the random vectors are independent and equally divided. The training set and random vectors are used to generate decision trees and a classifier would be created, where one of them is an input vector. After multiple decision trees are generated, the output category is determined by the mode of individual trees' category output, and one of them is selected [1].

In addition to high accuracy, random forest can process very high-dimensional data (multi-special micro data), without the need for special micro selection, and it is fast when training is completed. After the process, random forest can note which specific characteristics is more important, and be able to detect the influence between features. For imbalanced datasets, random forests can balance the error. If some of the features are missing, the accuracy can still be maintained, so it is the most popular among everyone. The random forest algorithm is used to model phishing website features as high as 97.7606%, as shown in Figure 7.



```
=== Summary ===

Correctly Classified Instances        2401               97.7606 %
Incorrectly Classified Instances        55                2.2394 %
Kappa statistic                          0.9547
Mean absolute error                      0.0616
Root mean squared error                  0.1423
Relative absolute error                 12.4622 %
Root relative squared error             28.6364 %
Total Number of Instances             2456

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.974     0.020     0.975     0.974     0.975     0.955     0.997     0.996     1
                 0.980     0.026     0.979     0.980     0.980     0.955     0.997     0.997     -1
Weighted Avg.    0.978     0.023     0.978     0.978     0.978     0.955     0.997     0.997

=== Confusion Matrix ===

    a    b   <-- classified as
 1066   28 |    a = 1
   27 1335 |    b = -1
```

Figure 7: Modeling the website dataset with random forests

3.3. Experimental Design

This article uses the websites of PhishTank, OpenPhish, and PhishRepo, all are well-known public information platform on phishing websites, to search for a total of 100 groups of short URLs that were confirmed as phishing websites from February to March 2022 (due to malicious short URLs' average survival time is 50 days, so only two months of the year's data are used for the experiment) and 100 legitimate websites with shortened URLs using URL shortening service platform. Through the mechanism of this experiment, the collected 200 sets of short URLs are used to compare the success rates of using three different machine learning algorithms to successfully identify legitimate websites and malicious phishing websites, and present the experimental results through a confusion matrix.

Confusion matrix is a visualization tool, often used in supervised learning (a type of machine learning). Each column of the matrix represents an instance prediction of a class, and each row represents an instance of an actual class. Through this matrix, it is easy to see whether the two different classes are confused, so it is called a confusion matrix, and the confusion matrix is composed of four elements: TP, TN, FP, and FN.

- TP (True Positive):
  The model prediction is true, and the real situation is also true.
- TN (True Negative):
  The model prediction is false, and the real situation is also false.
- FP (False Positive):
  The model prediction is true, but the real situation is false.
- FN (False Negative):
  The model prediction is false, but the real situation is true.

Table 2: Confusion matrix.

	Prediction is Positive	**Prediction is Negative**
Real Situation is Positive	True Positive (TP)	False Negative (FN)
Real Situation is Negative	False Positive (FP)	True Negative (TN)

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this study, 200 groups of short URLs (including 100 groups of malicious short URLs) were preprocessed using methods outlined in the first and second steps of the experimental process of this paper. After restoring them to the original URLs, they were compared with the database of the public phishing website platform, and 34 original URLs have been determined as reported phishing URLs. It can be easily seen as an act of phishing, as people intentionally packaged the original URLs into short URLs to deceive users. Not only this could evade the detection of antivirus software, but also because there is no longer the need to create a new phishing website and using the short URL service to redirect users is sufficient, saving a significant amount of cost. The remaining 66 URLs could proceed to the next step for comparison with the other 100 combined URLs who are not yet reported as malicious websites.

According to the aforementioned three machine learning algorithms, we present the experimental results in the form of confusion matrix as follows:

Table 3: Results predicted by Naïve Bayes model.

	Predicted as Phishing Site	**Predicted as Legitimate Site**
Truly a Phishing Site	TP：76	FN：24
Truly a Legitimate Site	FP：9	TN：91

Evaluate metrics with the Naïve Bayes model:

- Accuracy rate: (76+91)/200=83.5%
- Precision: 76/(76+9)=89.4%
- Recall rate: 76/(76+24)=76%

Table 4: Results predicted by the J48 model.

	Predicted as Phishing Site	**Predicted as Legitimate Site**
Truly a Phishing Site	TP：79	FN：21
Truly a Legitimate Site	FP：7	TN：93

Evaluate metrics with the J48 model:
- Accuracy rate: (79+93)/200=86%
- Precision: 79/(79+4)=95.18%
- Recall rate: 79/(79+21)=79%

Table 5: Results predicted by Random Forest model.

	Predicted as Phishing Site	**Predicted as Legitimate Site**
Truly a Phishing Site	TP：86	FN：14
Truly a Legitimate Site	FP：6	TN：94

Evaluate metrics with Random Forest model:
- Accuracy rate: (86+94)/200=90%
- Accuracy: 86/(86+4)=93.47%
- Recall rate: 86/(86+14)=86%

It can be seen from the above experimental data that the prediction accuracy of the data using the machine learning method is quite high, and the prediction result of the Random forest algorithm is the best, mainly because it is composed of different decision trees. , each decision tree is independent of each other and does not affect each other; and Naïve Bayes is lower in comparison, which can be attributed to this algorithm is an algorithm based on probability, predicting through the probability that may occur under certain conditions , so it is not so detailed, but it is relatively simple and saves time; the J48 algorithm between the two is more eclectic, unlike the Naïve Bayes algorithm that is too simple, but can prune the branches and leaves of the decision tree by itself, It is more flexible to use for processing unknown characteristics (attribute data).

V. CONCLUSIONS AND POSSIBLE FURTHER RESEARCH

The URL shortening service has become an indispensable tool for Internet users. Through it, a long URL can be shortened into a fixed-length URL, which is convenient for users to read and share. However, the convenience brought by it has also been maliciously used. Even under the protection of antivirus software, phishing URLs cannot easily be identified by users, which may lead to users being deceived by phishing websites. As a result, people's personal identity information and bank account passwords are invisibly defrauded, causing financial losses. Sometimes, a mobile phone or computer has become a part of this scheme. Through the mechanism proposed in this article, machine learning is used to verify the security of short URLs, analyze the effects of the three algorithms, and provide readers with a secure mechanism for verifying short URLs. Because the phishing website itself has a limited survival time, and will continue to introduce new ones in response to system vulnerabilities or the reduction of user security awareness, the phishing website database used in this study was established in 2015, and some of the extracted characteristics are no longer in line with the current situation. In the future, we will optimize the feature values of phishing websites, and we hope to identify phishing websites more effectively. At the same time, we also call for a service platform that shortens the URL to take social responsibility and firstly verify the original URL provided by the user, and if it belongs to malicious phishing websites, redirection services should not be provided, effectively reducing the risk of users being deceived.

REFERENCES

[1]. Daniel Chen, Using Random Forest Algorithm and Gradient Boosting to Analyze the Trading Strategies of Taiwan Fifty Corporation.
[2]. Min-Shiang Chiang, A Study on the Security of Short URLs, unpublished thesis.
[3]. Dayu Kao, Using Machine Learning Algorithms to Explore the Feature Values of Phishing Websites, unpublished thesis.
[4]. APWG, "Phishing Activity Trends Report 4th Quarter 2021", 2021, pp: 4.

[5]. Daejin Choi, Jinyoung Han, Selin Chun, Efstratios Rappos, Stephan Robert, Ted Taekyoung Kwon, "Bit.ly/practice: Uncovering content publishing and sharing through URL shortening services", Telematics and Informatics 35, 2018, pp: 1312.
[6]. Rana, Vijay. "Web URLs retrieval with least execution time using MPV clustering approach." International Journal of Information Technology (2020): 1-9.
[7]. Salo, Fadi, et al. "Data mining techniques in intrusion detection systems: A systematic literature review." IEEE Access 6 (2018): 56046-56058.
[8]. Martin Georgiev and Vitaly Shmatikov, "Gone in Six Characters: Short URLs Considered Harmful for Cloud Services", 2016, pp: 2.
[9]. Neha Gupta, Ponnurangam Kumaraguru, Anupama Aggarwal, "bit.ly/malicious: Deep Dive into Short URL based e-Crime Detection", June 2014.
[10]. Rana, Vijay. "A Web Extraction Browsing Scheme for Time-Critical Specific URLs Fetching." Proceedings of ICRIC 2019. Springer, Cham, 2020. 617-626.
[11]. Sarfaraz, Aaliya, and Ahmed Khan. "Feature selection based correlation attack on HTTPS secure searching." Wireless Personal Communications 103.4 (2018): 2995-3008.