International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 07 | July -2020 ISSN: 2582-3930

DETECTION OF MALICIOUS URL USING ENHANCED MACHINE LEARNING ALGORITHMS

S.Pavisya dharshini[$], M.Revathi[$], R.Suvetha[$],Mrs.K.V.Kiruthikaa*

[$]*Student, *Assistant Professor, Department of Computer Science and Engineering*
Bannari Amman Institute of Technology
Sathyamangalam, Erode, Tamilnadu, India

Abstract

The Primitive usage of URL is to use as a web address. Some URLs can be used to host unsolicited content that can potentially result in cyber attacks. These URLs are called malicious URLs. The inability of the end user system to detect and remove the malicious URLs can put the legitimate user in vulnerable condition. Furthermore, usage of malicious URLs may lead to illegitimate access to the user data by adversary. The main motive for malicious URL detection is that they provide an attack surface to the adversary. Malicious URL detector is a system which is mainly proposed to eliminate unwanted websites which affects online privacy and system health. Nowadays the world is revolving with full of internet activites so the dark web activites also increased which may be more harmful for today's youngsters and children which may affect their life cycle and privacy so in order to prevent these dark web activities we propose a solution with high integrated classification model by gathering web parameters and analysing it with dataset model. The system takes the web link as a input and scans for any malicious or malware contents inside it and alerts the user.

Keywords: Support Vector Machine, Malicious, Benign, classification, URL.

***Corresponding Author E-mail:suvetha.cs*16@bitsathy.ac.in**

I. INTRODUCTION

The advent of new communication technologies has had tremendous impact in the growth of business spanning across many applications including online-banking,e-commerce,and socialnetworking.In fact,in today's age it is almost mandatory tohave an online presence to run a successful venture.As a resul,the importance of the World WideWebhascontinuouslybeen increasing.Unfortunately,the technological advancements come coupled with new sophisticated techniques to attack and scam users.Such attacks include rogue websites that sell counterfeit goods,financial fraud by tricking users into revealing sensitive information which eventually lead to theft of money or identity,or even installing malware in the user's system.

There are a wide variety of techniques to implement such attacks, such as explicit hacking attempts, drive-by download, social engineering, phishing, watering hole, man-in-the middle, SQL injections, loss/theft of devices, denial of service, distributed denial of service, and many others.

Considering the variety of attacks, potentially new attack types, and the innumerable contexts in which such attacks can appear, it is hard to design robust systems to detect cyber- security breaches. The limitations of traditional security management technologies are becoming more and more serious given this exponential growth of new security threats, rapid changes of new IT technologies, and significant shortage of security professionals.

A URL has two main components: (i) protocol identifier (indicates what protocol to use) (ii) resource name (specifies the IP address or the domain name where the resource is located). The protocol identifier and the resource name are separated by a colon and two forwardslashes.Attackers use many other techniques to evade blacklist including: fast-flux, in which proxies are automatically generated to host the web page ;Algorithmic generation of new URLs.Additionally attackers can simultaneously launch more than one attack to alter the attack- signature making it undetectable by tools that focus on specific signatures. Blacklisting methods thus have severe limitations and it appears almost trival to bypass them, especially because blacklist are useless for making predictions on new URLs. To overcome these issues, in the last decade, researchers have applied machine learning techniques for Malicious URL Detection.

Machine Learning approaches, use a set of URLs as training data, and based on the statistical properties, learn a prediction function to classify a URL as malicious or benign. This gives them the ability to generalize to new URLs unlike blacklisting methods. The primary requirement for training a machine learning model is the presence of training data. In the context of malicious URL detection, this would correspond to a set of large number of URLs. Machine learning can broadly be classified into supervised, unsupervised, and semi- supervised, which correspond to having the labels for the training data, not having the labels, and having labels for limited fraction of training data, respectively. Labels correspond to the knowledge that a URL is malicious or benign.

II. LITERATURE SURVEY

Wei Zhang and Huan Ren 2016 proposed a novel anti-phishing framework that employs feature engineering including feature selection and feature extraction. First, perform feature selection based on genetic algorithm (GA) to divide features into critical features and non-critical features. Then the non-critical features are projected to a new feature by implementing feature extraction based on a two-stage projection pursuit (PP) algorithm. Finally, the critical features and the new feature are taken as input data to construct the detection model. The anti-phishing framework does not simply eliminate the non-critical features, but considers utilizing their projection in the process of classification, which is different from literatures. Experimental results show proposed framework is effective in detecting phishing webpages.

Roshani.k and Chaudhari 2016 stated that malicious tweets having URLs for spam distribution. Conventional twitter spam detection methods, take advantages of accounts features such as the ration of tweets containing URLs and the date of creating an account or relation features in the twitter graph. These detection methods are ineffective against feature fabrications are consume much time and resource. The proposed system to find malicious url and spam and to identify whether a given tweet is spam or not in social network such as facebook and twitter.

By collecting dataset and training the classifier the input tweet. The naive bayes algorithm a supervised learning model with associated learning algorithms which are used to analyze data used for classification and regression analysis. After the classification the sensitivity of each tweet is calculated. Then the experimental results are found that the trained classifier is shown to be accurate and has low false positive and negatives.

Guolin Tan et.al in 2018 stated that Hackers can implant the malwares like Trojan, Worms, Spam etc in the webpages to steal information and acquire money illegally. In fact it is noted that close to one third of all websites are potentially malicious in nature. Therefore, it makes sense to quickly detect malicious URLs on the internet different from most of previous methods. The proposed method for online malicious Url detection based on adaptive learning. By collecting the networks, the machine learning models to train inorder to detect malicious URLs, but there is a serious problem in dynamically changing environments where the statistical properties of target variable change over time which is known as concept drift. To address this problem, a non parametric test are apply to correctly detect concepts drifts in adaptive learning. Extensive experiments with different types of concept drifts are performed to demonstrate the feasibility of proposed method on both artificial and real datasets. The empirical study shows that this approach has good performance in detecting malicious URLs and concept drifts.

Manamohana K et.al 2019 stated that malicious urls are one of the ways to hack a user's personal information. Inorder to overcome this limitations posed by the primitive classification methodologies like Black-Listing, Heuristic classification Research has been carried over the several areas and machine learning is one of the promising approaches to effectively classify the URLs explains one of the several ways to leverage the machine learning in URL detection. Using the Supervised Machine Learning concepts such as Random Forest Model can classify at 89% without any tuning and feature selection. Some Approaches in which detailed feature extraction required for precision classification. Uses the word level and character level .

Convolutional Neural Networks, as the underlying neural networks are quite handy in dealing with image data for computer vision tasks especially in deriving and learning from the salient features of the images from the raw pixel values.This approach produced better results by classifying the URL at a precision of 94%. This methodology uses the URL detection at the Character level and word level and finally to complete URL. Usually, the problem will arise during the gathering of the data uses the manual updates of the URLs which is so hectic in reality, the general principle is to make the automated model in order to collect the data, but they are so many difficulties in reality.

Doyen Sahoo et.al 2019 implemented an intelligent system to detect malicious url by converting URLs into feature vectors, many of these learning algorithms can be generally applied to train a predictive model in a fairly straightforward manner. However,to effectively solve the problem, some efforts have also been explored in devising specific learning algorithms that either exploit the properties exhibited by the training data of Malicious Url or address some specific challenges which the application faces. The learning algorithms that have been applied for this task, and also suggest suitable machine technologies that can be used to solve specific challenges encountered. Batch Learning algorithms work under the assumption that the entire training data is available prior to the training task. Online learning algorithms treat the data as a stream of instances and learn a prediction model by sequentially making prediction and updates. This makes them extremely scalable compared to batch algorithms. The extensions of Online Learning to cost-sensitive and active learning scenarios. Representation learning methods, which are further categorized into Deep Learning and Feature Selection techniques.algorithms, in challenges specific to Malicious URL Detection are addressed, including unsupervised learning,similarity learning and string pattern Matching.

III MACHINE LEARNING ALGORITHM

Use Machine Learning Technology for two key purposes identify important insights in data, and prevent fraud. The insights can identify investment opportunities, or help investors know when to trade. It is the ability for computers to learn and act without being explicitly programmed. Machine learning focuses on the development of computer programs That can access data and use it learn for themselves.

Design and propose a Malicious URLs detection system using machine learning techniques.These approaches try to analyze the information of URL and its corresponding websites or webpages,by extracting good feature representati--on of URLs and training a prediction model on training data of both malicioUos and benign URLs.There are two types of features that can be used static and dynamic features.In this project we used two machine learning algorithms support vector machine(SVM) and Random Forest classifier.

TYPES OF MACHINE LEARNING

There are some deviation of how to describe the types of Machine Learning Algorithms but it regularly they can be separated into categories according to their cause and the major categories are the following: Supervised learning. Unsupervised Learning, Semi-supervised learning.

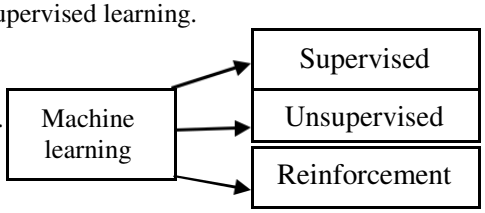
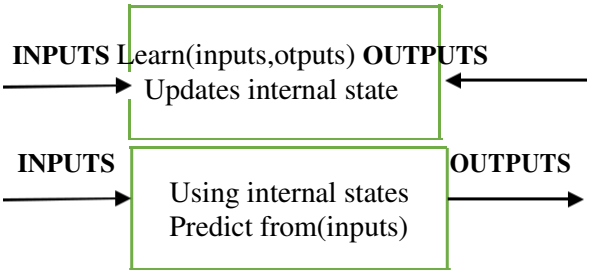


Figure 3.1 Types of machine learning

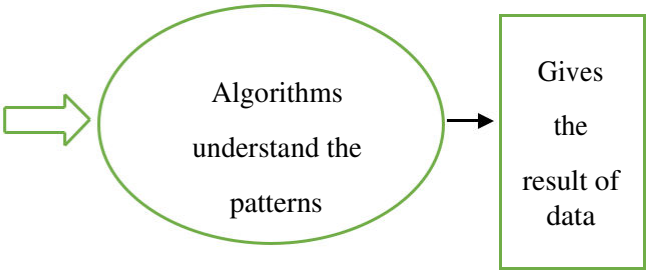
SUPERVISED LEARNING

Data can be gathered from several sources and information. In supervised learning, input variables sometimes called as as Independent variables and Output variables also known as Targets, Dependent variables, Labels. Supervised learning with the idea of function approximation, where fundamentally training an algorithm and in the end of the method chooses the function that most excellently describes the input data, the one that for a given X makes the best.



UNSUPERVISED LEARNING

In Unsupervised Learning output variable date to predict is not available.With only inputs and algorithm,it then identify what the data is and group or clusters the data .Being as a human ,using the unsupervised data,it does not recognize what the data is,the data is completely unstructured.



REINFORCEMENT LEARNING

It is most advanced in above types of Machine learning. Reinforcement learning occurs when you present the algorithm with example that not have labels, as in unsupervised learning. With an instance of positive or negative feedback according to the solution that the algorithm proposes. Reinforcement learning is linked to applications for which the algorithm must create decisions and the decisions stand consequences.

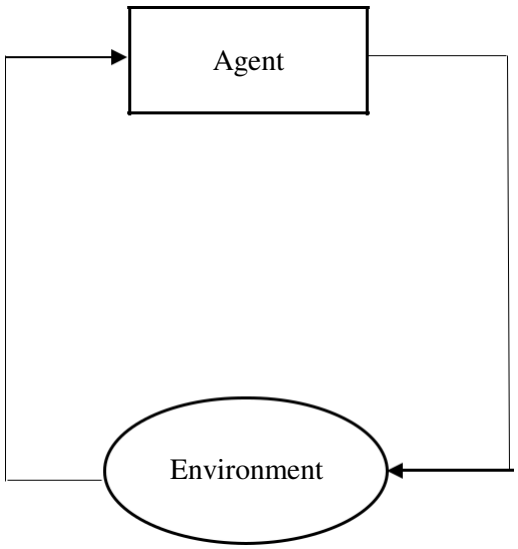


Figure 3.4 Reinforcement Learning

PHASES OF MACHINE

LEARNING PRE-PROCESSING

Pre-processing refers to the transformation our data before feeding it to the algorithm. Data pre-processing is a method that is used to exchange the unprocessed data into a clean data set.

In other word, she data is gathered from various sources it is collected in raw format which is not possible fortheanalysis.In Real-world data is often imperfect, inconsistent, and missing in definite behaviors and is likely to include many errors. Data preprocessing is a verified method of resolve such problems.

FEATURE EXTRACTION

Feature extraction is a conversion of raw data into features appropriate for modeling which is used in supervised learning. Feature extraction can extract text from n-grams, images from CNN‘s texts and date and time from month year week.

FEATURE SELECTION

Feature selection is also known as attribute selection. It is an automatic selection of variables in your data like columns in the table that are most applicable to the analytical modeling. Feature selection is different from dimensionality reduction.

Both methods try to reduce the number of variables in the dataset, but in dimensionality reduction and introducing new group of attributes. This technique is used for selecting the features that explain the most of the target variable.

CLASSIFICATION

Classification is the technique of predicting the class of the given data. Classes are also called as labels. Classification modeling is the task of predicting approximating a mapping function(z) from input variables(a) to discrete output variables (b).Classification is considered an example of supervised learning that learning where the training set of data of correctly identified and the observations is available. The unsupervised procedure is known as clustering, and involves combining data into types based on some measure of some similarity or distance.

IV. EXISITING SYSTEM

The Existing system uses Batch learning, Decision Tree and Navie Bayes to detect the Malicious Urls.Machine learning technique typically comprises of two steps: one is to obtain the appropriate feature representation that it could provide the determining insights in finding the Malicious URLs, and the second is to use this representation to train a learning-based prediction mechanism. A novel classification method to address the challenges faced by the traditional mechanisms in malicious URL detection. The proposed classification model is built on sophisticated machine learning methods that not only takes care about the syntactical nature of the URL,but also by batch learning. however, they cannot cope up with the evolving attack techniques. Recent statistics imply that there is 20 - 25% growth in the attacks yearly and the threats that are coming from the newly created URLs are on the rise. One serious limitation of these techniques is that they are inefficient to classify the newly generated URLs.

V. PROPOSED SYSYTEM

In Today’s World it is almost mandatory to have an online presence to run a successful venture. As a result, the importance of the World Wide Web has continuously been increasing. Unfortunately, the technological advancements come coupled with new sophisticated techniques to attack and scam users. The main motive for malicious URL detection is that they provide an attack surface to the adversary. Malicious URL detector is a system which is mainly proposed to eliminate unwanted websites which affects online privacy and system health. The proposed solution which is high integrated classification model by gathering web parameters and analysing it with dataset model. The system takes the web link as a input and scans for any malicious or malware contents inside it and alerts the user.

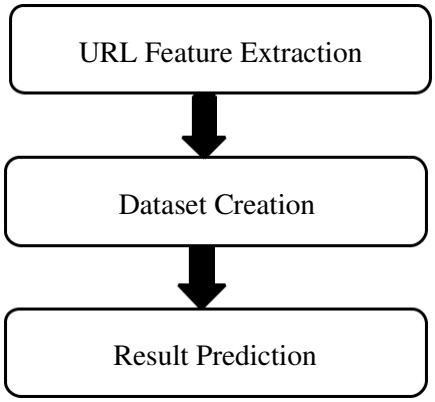


Figure 5.1 Flow chart for Proposed System

The Machine learning algorithms used in the project are Support Vector Machine(SVM) and Random Forest Classifier.

SUPPORT VECTOR MACHINE

Support Vector Machines (SVM) are another powerful supervised learning technique that can be used for classification and regression analysis. The base idea of SVM training algorithm starts from a given set of training examples (support vectors), where each one is marked for belonging to one of two categories, as a non-probabilistic binary linear classifier. Then, the SVM can be seen as a representation of the examples as points in space, with the goal of separating examples in categories divided by a clear gap. More formally, SVM tries to find a hyperplane or set of hyperplanes in a high-or infinite-dimensional space, such as a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class, minimizing the generalization error of the classifier. Unclassified instances then are mapped into the same space for classification.

The hyperplanes in the higher-dimensional space are defined as the set of points whose dot product with a vector in that space is constant. The vectors defining the hyperplanes can be chosen to be linear combinations of images of feature vectors. Several extensions of SVM such as support vector clustering or transductive Support Vector Machines are worthnoting.SVM remains as a competitive performing machine learning technique for supervised classification.

RANDOM FOREST CLASSIFIER

Random Forest (RF) is a well-known ensemble learning method for supervised classification or regression. This machine learning technique operates by building an ensemble of random decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Therefore a RF is a classifier consisting in a collection of tree structured classifiers which uses random selection in two moments. .

In a first step, the algorithm selects several bootstrap samples from the historical data. For each bootstrap selection , the size of the selected data is roughly 2/3rd of the total training data. Cases are selected randomly with replacement from the original data and observations in the original data set that do not occur in a bootstrap sample are called out-of-bag (OOB) observation. In a second step, a classification tree is trained using each bootstrap sample, but only a small number of randomly selected variables (commonly the square root of the number of variables) are used for partitioning the tree. The OOB error rate is computed for each tree, using the rest (36.8%) of historical data. The overall OOB error rate is then aggregated, observe that RF does not require a split sampling method to assess accuracy of the model. The final output of the model is the mode (or mean) of the predictions from each individual tree. Random Forest comes at the expense of some loss of interpretability, but generally greatly boosts the performance of the final model, becoming one of the most likely to be the best performing classifier in real-world classification problem.

VI. IMPLEMENTATION

In order to identify these malicious sites, several studies in the literature hassle this issue from a Machine Learning standpoint. That is, they compile a list of URLs that have been categorized as both malicious and benign and characterize every URL via a set of attributes. Classification algorithms are then predicted to take a look at the boundary among the decision classes. Malicious URLs for the experimentation were accumulated from Kaggle, which is an open source community site wherein manually verified Malicious URLs are uploaded. Dataset of this project is taken from this Open Source.

Numerous researchers have used one kind of sets of features to solve the problem of URL classification.Even though interested in machine learning techniques, but out of all Support Vector Machine(SVM) and Random Forest classifier provided the better results this is because of the effective learning rate and quite suitable for the feature extraction. Then, create a Graphical User Interface which is used to get the URL from user to check whether it is benign or malicious, For that Tkinter is used which is the standard GUI library for Python.

Python when combined with Tkinter provides a fast and easy way to create GUI applications. Tkinter provides a powerful object-oriented interface to the Tk GUI toolkit. Then proposed a solution with high integrated classification model by gathering web parameters and analyzing it with dataset model and predict the result. Based on the result it identifies whether the URL is Benign or malicious or malware.

METHODOLOGY

Feature Extraction - GUI Tkinter development,Collection of labelled URL data
Dataset Extraction –Model Training, Batch Training
Binary Classification Using Support Vector Machine and Random Forest
Result Prediction

GUI TKINTER DEVELOPEMENT

Tkinter is the standard GUI library for Python. Python when combined with Tkinter provides a fast and easy way to create GUI applications. Tkinter provides a powerful object-oriented interface to the TkGUI toolkit. Tkinter provides various controls, such as buttons, labels and text boxes used in a GUI application. These controls are commonly called widgets.Graphical User Interface is created which is used to get the URL from user to check whether it is benign or malicious.

COLLECTION OF LABELLED URL

Kaggle is the open source that allow the registered users to add new malicious URLs that are not in the existing one.It is a platform for predictive modelling and analytics competitions in which companies and researchers post data and statisticians and data miners compete to produce the best models for predicting and describing the data. The Dataset used here is taken from the Open Source kaggle**.**

MODEL TRAINING

Training a model simply means learning (determining) good values for all the weights and the bias from labeled examples. In supervised learning, a machine learning algorithm builds a model by examining many examples and attempting to find a model that minimizes loss; this process is called risk minimization.

BATCH TRAINING

Batch means a group of training samples. In gradient descent algorithms, it calculate the sum of gradients with respect to several examples and then update the parameters using this cumulative gradient.

RESULT PRDICTION

Crawling URL feed from existingdatabase and capturing live URL features and comparing with training dataset to predict the result**.**

VII. SYSTEM REQUIREMENTS

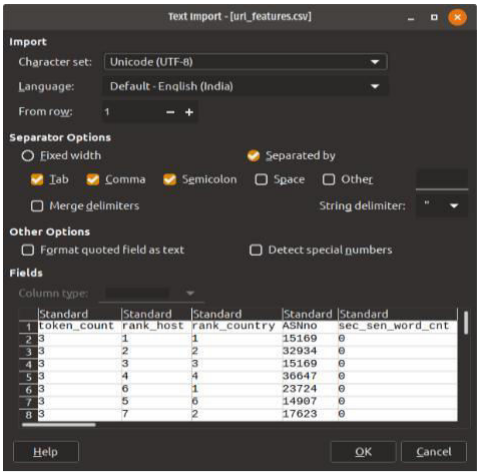
HARDWARE REQUIREMENTS

GB RAM
1.2GHZ
INTEL QUADCORE

SOFTWARE REQUIREMENTS

LINUX(Ubuntu 18.04.3)
Geany
Python 2.7
Pandas

VIII. RESULTS AND DISCUSSION NORMALIZED DATA



	Standard token_count	Standard rank_host	Standard rank_country	Standard ASNno	Standard sec_sen_word_cnt
1	3	1	1	15169	0
2	3	2	2	32934	0
3	3	3	3	15169	0
4	3	4	4	36647	0
5	3	6	1	23724	0
6	3	5	6	14907	0
7	3	7	2	17623	0

Figure 8.1 Normalized Data

PREDICTION OF MALWARE

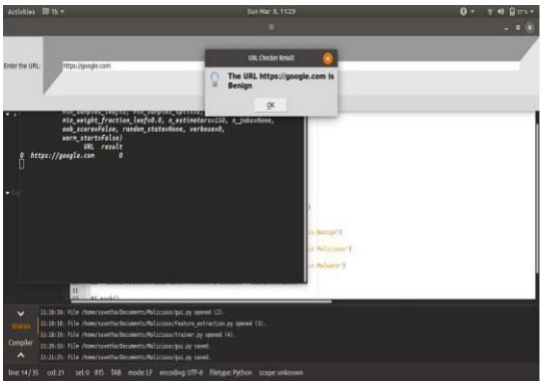


Figure 8.2 Prediction of Benign URL

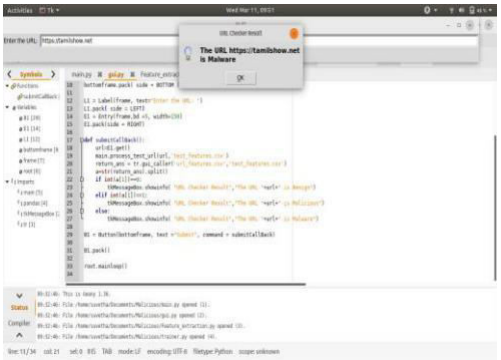


Figure 8.3 Prediction of Malware URL

IX. CONCLUSION AND FUTURE WORK

The aim of the project is to described how a machine can able to judge the URLs based upon the given feature set.The proposed system is designed in such a way with high integrated classification model by gathering web parameters and analyzing it with dataset model .The system takes the web link as the input and scan for benign or malicious or malware content inside it and alert the user. When traditional method fall short in detecting the new malicious URLs on its own, our proposed method can be augmented with it and is expected to provide improved results. The future work is to fine tuning the machine learning algorithm that will produce the better result by utilizing the given feature set and block the malicious websites.

REFERENCES

[1] G. A. Montazer, S. Yarmohammadi, "Detection of phishing attacks in Iranian e-banking using a fuzzy–rough hybrid system," Applied Soft Computing, Vol. 35, pp. 482-492,2015.

[2] Hinton, G. E., Osindero, S., and Teh, Y., "A fast learning algorithm for deep belief nets", Neural Computation, 18 (7):1527–1554,2006.

[3] RoshaniK. Chaudhari, D. M. Dakhane, "Machine Learning Approach for Detection of Malicious Urls and Spam in Social Network ", International Research Journal of Engineering and Technology (IRJET), vol. 03, no. 05, pp. 2395-0072 ,2016.

[4] Haijun Zhang, Gang Liu, Tommy WS Chow, and Wenyin Liu. 2011. Textual and visual content-based anti-phishing: a Bayesian approach. IEEE Transactions on Neural Networks (2011).

[5] Wen Zhang, Yu-Xin Ding, Yan Tang, and Bin Zhao. 2011. Malicious web page detection based on on-linelearningalgorithm. In Machine Learning and Cybernetics (ICMLC), 2011 International Conference on. IEEE.

[6] Wei Zhang, REN Huan, and Qingshan Jiang. 2016. Application of Feature Engineering for Phishing Detection. IEICE TRANSACTIONS on Information and Systems(2016).