International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 11 – No. 11, March 2017 – www.ijais.org

Efficient Detection of Legitimate and Malicious URLs using ID3 Algorithm

Yogesh Dubey
Student
St. John College Of
Engineering and Management
Palghar

Pranil Chaudhari
Student
St. John College Of
Engineering and Management
Palghar

Shaldon Chaphya
Student
St. John College Of
Engineering and Management
Palghar

Tina D’abreo
Lecturer
Computer Department
St. John College Of
Engineering and Management
Palghar

ABSTRACT

Malicious websites are one of the serious threat over the internet. Ever since the inception of the internet, there has been a rise in malicious content over the web such has terrorism, financial fraud, phishing and hacking that targets user’s personal information. Till date, the various systems have been used for the detection of a malicious website based on text and content of the websites. This method has some disadvantages and the numbers of victims have therefore continued to increase. Here we developed a system which helps the user to identify whether the website is malicious or not. Our system identifies whether the site is malicious or not through URL. The proposed system is fast and more accurate compared to current system. The classifier is trained with datasets of 1000 malicious sites and 1000 legitimate site URLs. Trained classifier is used for detection of malicious URLs.

Keywords

Malicious URLs, Classifier, Feature Extraction, ID3 Algorithm.

1. INTRODUCTION

Since the growth of the internet environment, there has been a rise in malicious content over the internet get user’s confidential information or spreading of false information. Similarly, the quality and quantity of web attacks have increased. There is no doubt that harmful websites can be extremely damaging all organizations and access user’s information in order to gain access to that business network.

The number of attacks has increased three to five-fold in less than 5 years, resulted loss is billions and the number of the malicious websites is raising day by day. The current system which is used for detection of malicious websites is not effective in detecting the temporary or new malicious websites. In this paper, we propose a system which uses an automated classifier for the detection of malicious websites using URL features.

2. LITERATURE SURVEY

In the paper [1], the authors have described URL is used for detection of malicious websites. The focus of classification of URLs is based on host properties and a bag of words. The motivation is to provide inherently better coverage than blacklisting based approaches while avoiding the client-side over-head and risk of approaches that analyze web content. The detection of a malicious website based on only host-based properties is not reliable.

In paper [2], the heuristic based approach is used for the detection of phishing websites. The phishing websites are identified based on features of the URLs. The features of URL are used for the phishing site detection.

In paper [3], the characteristic of websites is used for checking the trustworthiness, the filtering of website trustworthiness is based on five major areas as Authority, Related resources, Popularity, Age, hits, and Recommendation. The website trustworthiness is calculated based on these eighteen factors of each URL and it is stored thereby increasing the performance in retrieving the trustworthy websites.

In the paper [4] author describes various features of URL are extracted and analyzed based on the feature selection methods and classification algorithm for phishing websites detection.

In paper [5] author has proposed an URL based method for the identification of phishing websites through URL. The proposed system mainly focuses similarities between the legitimate website and phishing website. The ranking of phishing website is also considered. The system effectiveness is limited to detection of phishing websites only.

3. SYSTEM ARCHITECTURE

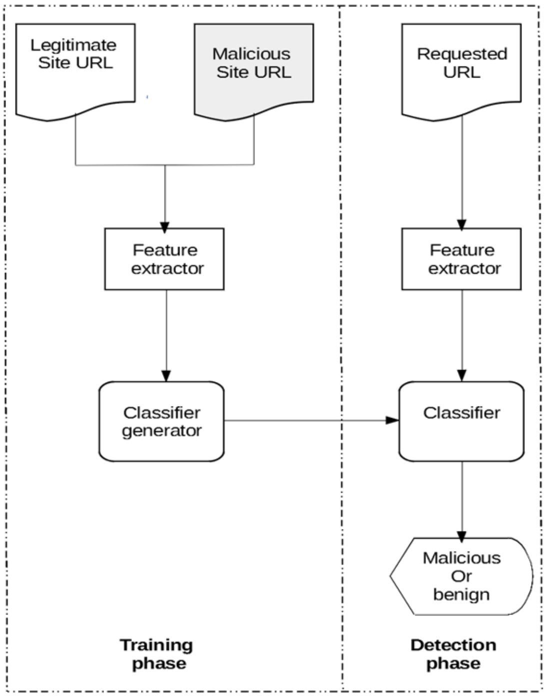


Figure 1.Architecture Diagram [2]

3.1 Training Phase

In the training phase, a classifier is generated using URLs of malicious sites and legitimate sites collected in advance. The collected URLs are transmitted to the feature extractor, which extracts feature values through the predefined URL-based features. The extracted features are stored as input and passed to the classifier generator, which generates a classifier by using the input features and the machine learning algorithm.

3.2 Detection Phase

In the detection phase, the classifier determines whether a requested site is a malicious site. When a page request occurs, the URL of the requested site is transmitted to the feature extractor, which extracts the feature values through the predefined URL-based features. Those feature values are inputted to the classifier. The classifier determines whether a new site is a malicious site based on learned information. It then alerts the page-requesting user about the classification result.

3.3 Feature Extractor

The feature extractor extracts the features from the URL. The proposed system extract following certain features of the URL stated below:

1. Length of URL: The length of URL is calculated and on the basis of length it is determined whether the URL can be malicious or not.

2. Top-level domain: The number of top level domain or position of top level domain in the URL.

3. Suspicious character: If there is any suspicious character in the URL like”*”.

4. Protocol: Type of protocol URL has like HTTP or https.

5. The length of the subdomain: The length of the subdomain is calculated and whether it is too long or short is determined.

6. A number of the subdomain: the number of subdomains present in the URL.

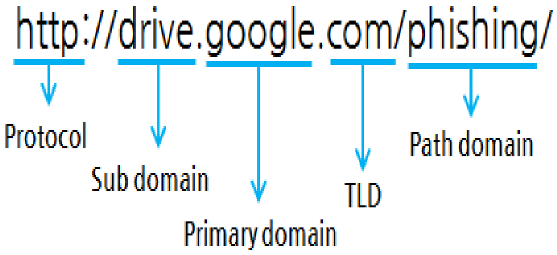


Figure 2. The structure of URL [10]

3.4 Classifier

In the training phase, the classifier is generated by training it with datasets of malicious and legitimate website URLs. The feature extractor is used to extract the mentioned features of the URL. Decision tree algorithm is used for generating a classifier. The generated classifier is further used in detection phase where the user will enter the URL to check the whether the URL is malicious or not.

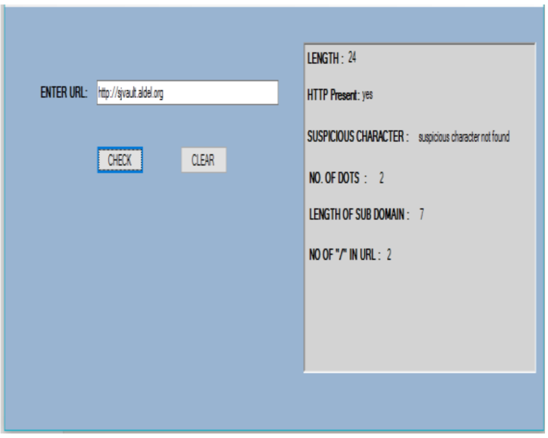
4. EXPECTED OUTCOME

The whole system is divided into two phase. The first part consisting of a training phase and the second part is detection phase.

In the training phase, the classifier is generated with the help of dataset of the malicious and legitimate website. All the described features are extracted from the dataset are URLs and are passed to the id3 algorithm we used for making the decision tree which generates rules for the classifier and trains the classifier.

In the detection phase URLs is entered by the user. The feature extractor will extract the described features and its passed on the already trained classifier which is able to identify the whether the entered URL is malicious or legitimate.

5. RESULTS


ENTER URL: http://sjvault.aldel.org

LENGTH : 24

HTTP Present: yes

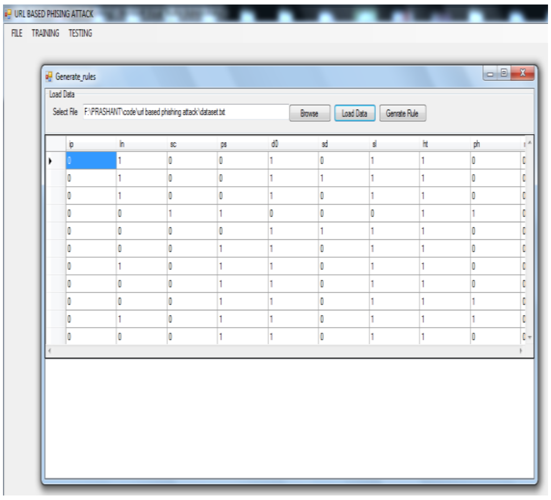
SUSPICIOUS CHARACTER : suspicious character not found

NO. OF DOTS : 2

LENGTH OF SUB DOMAIN : 7

NO OF "/" IN URL : 2

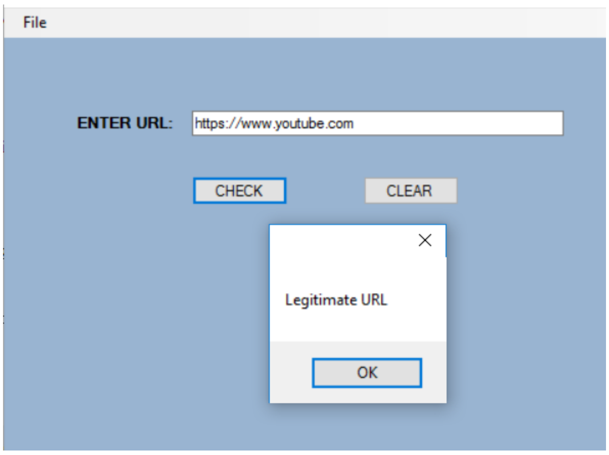
Figure 3: Feature Extractor



Select File F:\PRASHANT\code\url based phishing attack\dataset.txt

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

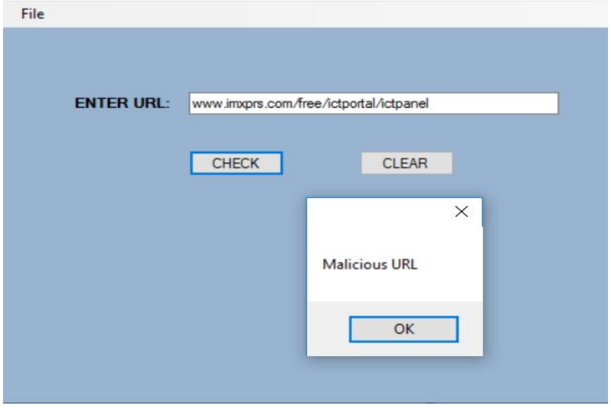
Figure 4: Loading Dataset



ENTER URL: https://www.youtube.com

Legitimate URL

Figure 5: Classifier Result 1



ENTER URL: www.imxprs.com/free/ictportal/ictpanel

Malicious URL

Figure 6: Classifier Result 2

6. CONCLUSION

We have developed a system for detection of malicious websites through URL which based on an automated classifier. The classifier is trained with the dataset of legitimate and malicious websites. The trained classifier is for the detection of any URL. Further, the accuracy of the system increases as the classifier is trained with more data set.

7. REFERENCES

[1] Ma, Justin, et al. "Beyond Blacklists: learning to detect malicious websites from suspicious URLs." Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009

[2] Jin-Lee Lee,Doung-Hyun Kim,Chang-Hoon Lee. “Heuristic-based Approach for Phishing Site Detection Using URL Features” Third Intl. Conf. on Advances in Computing, Electronics and Electrical Technology - CEET 2015.

[3] Sana Ansari and Jayant Gadge. “Architecture for Checking Trustworthiness of Websites “International journal of computer application, Volume 44, April 2012

[4] Mustafa Aydin and Nazife Baykal “Feature Extraction and Classification Phishing Websites Based on URL” Cyber Defence and Security Laboratory of METU-COMODO, IEEE CNS 2015.

[5] Nguyen, Luong Anh Tuan, et al. "A novel approach for phishing detection using URL-based heuristic." Computing, Management and Telecommunications (ComManTel), 2014 International Conference on. IEEE, 2014.

[6] Canali, Davide, et al. "Prophiler: a fast filter for the large-scale detection of malicious web pages." Proceedings of the 20th international conference on World wide web. ACM, 2011.

[7] Sumalatha Ramachandran, Sujaya Paulraj, Sharon Joseph and Vetriselvi Ramaraj, “Enhanced Trustworthy and High-Quality Information Retrieval System for Web Search Engines”, IJCSI International Journal of Computer Science Issues, Vol. 5, October 2009, pp38-42.

[8] https://en.wikipedia.org/wiki/Uniform_Resource_Locator

[9] https://www.phishtank.com/developer_info.php

[10] https://en.wikipedia.org/wiki/ID3_algorithm