MALICIOUS WEBSITE DETECTION

ONG VIENNA LEE

BACHELOR OF COMPUTER SCIENCE
(COMPUTER SYSTEM AND NETWORKING)

UNIVERSITI MALAYSIA PAHANG
SUPERVISOR’S DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Computer Science (Computer Systems and Networking) with Honours.

(Supervisor’s Signature)

Full Name : EN. MOHD FAIZAL AB RAZAK
Position : SENIOR LECTURER
Date : 10th JANUARY 2019
STUDENT’S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Universiti Malaysia Pahang or any other institutions.

____________________________
(Student’s Signature)

Full Name : ONG VIENNA LEE
ID Number : CA15050
Date : 10th JANUARY 2019
MALICIOUS WEBSITE DETECTION

ONG VIENNA LEE

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Computer Science (Computer Systems and Networking)

Faculty of Computer Systems & Software Engineering
UNIVERSITI MALAYSIA PAHANG

JANUARY 2019
ACKNOWLEDGEMENTS

I would like to take this opportunity to express my gratitude towards my supervisor En. Mohd Faizal Ab Razak who gave me guidance throughout the research. It is a great honor to be able to do this research under his supervision. He provided me with a lot of insights and advices all the time.

An incomparable appreciation towards my family and friends that have been supporting and giving me encouragement throughout the research. Their motivation played the role in keeping me to complete this thesis in time.
ABSTRAK

Malicious websites are among the major security threats on the Internet. This threat has been existing for years yet the best solution to overcome it has not been implemented by many people. Most of the existing methods for detecting malicious websites are focusing towards specific attacks. However, attacks are getting more complex and hackers have become more sophisticated with their blended techniques to evade existing countermeasures. In this thesis, a method will be introduced. With previous existing methods in consideration, the method to use for this project is by using heuristic-based detection with machine learning technique and the feature that will be used together with the technique is URL based feature. The purpose of this method is to classify benign and malicious website using machine learning and then will automatically detect malicious websites. By using this method is also to ensure the detection accuracy is high and all malicious websites can be detected even the latest one prompted by the hackers. In conclusion, the proposed method is the most effective way to detect malicious websites and easy to be implemented.
# TABLE OF CONTENT

DECLARATION

TITLE PAGE

ACKNOWLEDGEMENTS ii

ABSTRAK iii

ABSTRACT iv

TABLE OF CONTENT v

LIST OF TABLES viii

LIST OF FIGURES ix

LIST OF ABBREVIATIONS x

CHAPTER 1 INTRODUCTION 1

1.1 Introduction 1

1.2 Problem Statement 2

1.3 Objectives 3

1.4 Scope 3

1.5 Significance 3

1.6 Report 4

CHAPTER 2 LITERATURE REVIEW 6

2.1 Introduction 6

2.1.1 Signature-based Detection Technique 8

2.1.2 Heuristic-based Detection Technique 9

2.1.3 N-Gram Technique 10

2.2 Feature Selection 10
2.2.1 URL Feature

2.2.2 Host Based Feature

2.2.3 Content Based Feature

2.2.4 Graph Based Feature

2.2.5 Blacklist Feature

2.3 Comparison of Commonly-Used Method

2.4 Related Works on the Research

2.5 Conclusion

CHAPTER 3 METHODOLOGY

3.1 Introduction

3.2 Methodology

   3.2.1 Planning Phase

   3.2.2 Analysis Phase

   3.2.3 Design Phase

   3.2.4 Development Phase

   3.2.5 Testing phase

   3.2.6 Thesis Writing

3.3 Gantt Chart

3.4 Implementation

CHAPTER 4 IMPLEMENTATION, TESTING AND RESULT DISCUSSION

4.1 Introduction

4.2 Collect Dataset

4.3 Data Analysis

4.4 Feature Selection Results
4.4.1 Particle Swarm Optimization (PSO) 34
4.4.2 Information Gain 35

4.5 Algorithms Testing Results 36
4.5.1 Random Forests 36
4.5.2 Naïve Bayes 36
4.5.3 K-Nearest Neighbors (k-NN) 37
4.5.4 Support Vector Machines (SVM) 37
4.5.5 Adaptive Boost (AdaBoost) 37

4.6 Implementation of the tool 45
4.7 User Manual 46

CHAPTER 5 CONCLUSION 49

5.1 Introduction 49
5.2 Research Constraints 49
5.3 Achievement 50
5.4 Future Work 50

REFERENCES 51
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.1</td>
<td>Type of features in relation to type of attack</td>
<td>11</td>
</tr>
<tr>
<td>Table 2.2</td>
<td>Comparison of commonly used method in malicious website detection</td>
<td>15</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>Hardware requirements</td>
<td>23</td>
</tr>
<tr>
<td>Table 3.2</td>
<td>Software requirements</td>
<td>23</td>
</tr>
<tr>
<td>Table 3.3</td>
<td>Total number of samples</td>
<td>23</td>
</tr>
<tr>
<td>Table 4.1</td>
<td>Features of URLs</td>
<td>33</td>
</tr>
<tr>
<td>Table 4.2</td>
<td>The explanation for the evaluation</td>
<td>38</td>
</tr>
<tr>
<td>Table 4.3</td>
<td>Evaluation results</td>
<td>39</td>
</tr>
<tr>
<td>Table 4.4</td>
<td>The accuracy results comparison with past research papers</td>
<td>44</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.1</td>
<td>Summary of each chapter</td>
<td>4</td>
</tr>
<tr>
<td>Figure 2.1</td>
<td>Classification of most frequently exploited websites</td>
<td>8</td>
</tr>
<tr>
<td>Figure 2.2</td>
<td>Detection tool techniques</td>
<td>8</td>
</tr>
<tr>
<td>Figure 2.3</td>
<td>Feature selection types</td>
<td>11</td>
</tr>
<tr>
<td>Figure 2.4</td>
<td>The framework of the method</td>
<td>16</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>Waterfall model phase</td>
<td>21</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>Example of data visualization</td>
<td>24</td>
</tr>
<tr>
<td>Figure 3.3</td>
<td>General framework for malicious URL detection using machine learning</td>
<td>25</td>
</tr>
<tr>
<td>Figure 3.4</td>
<td>Flowchart for determining and detecting malicious and benign website</td>
<td>26</td>
</tr>
<tr>
<td>Figure 3.5</td>
<td>Gantt chart for project system development</td>
<td>29</td>
</tr>
<tr>
<td>Figure 4.1</td>
<td>The user interface of Weka</td>
<td>34</td>
</tr>
<tr>
<td>Figure 4.2</td>
<td>Output for Particle Swarm Optimization</td>
<td>35</td>
</tr>
<tr>
<td>Figure 4.3</td>
<td>Output for Info Gain</td>
<td>35</td>
</tr>
<tr>
<td>Figure 4.4</td>
<td>Formulas for evaluation results</td>
<td>40</td>
</tr>
<tr>
<td>Figure 4.5</td>
<td>TP Rate comparison among the algorithms</td>
<td>42</td>
</tr>
<tr>
<td>Figure 4.6</td>
<td>FP Rate comparison among the algorithms</td>
<td>42</td>
</tr>
<tr>
<td>Figure 4.7</td>
<td>Precision comparison among the algorithms</td>
<td>43</td>
</tr>
<tr>
<td>Figure 4.8</td>
<td>Comparison of accuracy results between the algorithms</td>
<td>43</td>
</tr>
<tr>
<td>Figure 4.9</td>
<td>The implementation process</td>
<td>45</td>
</tr>
<tr>
<td>Figure 4.10</td>
<td>Load Unpacked tab</td>
<td>46</td>
</tr>
<tr>
<td>Figure 4.11</td>
<td>‘ScripSafe’ file uploaded/installed</td>
<td>46</td>
</tr>
<tr>
<td>Figure 4.12</td>
<td>Example of the coding used</td>
<td>47</td>
</tr>
<tr>
<td>Figure 4.13</td>
<td>Blocked website due to its malicious features</td>
<td>47</td>
</tr>
<tr>
<td>Figure 4.14</td>
<td>The access for benign website</td>
<td>48</td>
</tr>
</tbody>
</table>
## LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbr</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQL</td>
<td>Structured Query Language</td>
</tr>
<tr>
<td>HTML</td>
<td>HyperText Markup Language</td>
</tr>
<tr>
<td>URL</td>
<td>Uniform Resource Locator</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>DNS</td>
<td>Domain Name Server</td>
</tr>
<tr>
<td>HTTP</td>
<td>HyperText Transfer Protocol</td>
</tr>
<tr>
<td>DHTML</td>
<td>Dynamic HyperText Markup Language</td>
</tr>
<tr>
<td>IP</td>
<td>Internet Protocol</td>
</tr>
<tr>
<td>DSL</td>
<td>Digital Subscriber Line</td>
</tr>
<tr>
<td>CSS</td>
<td>Cascading Style Sheets</td>
</tr>
<tr>
<td>I/O</td>
<td>Input/Output</td>
</tr>
<tr>
<td>SDLC</td>
<td>System Development Life Cycle</td>
</tr>
<tr>
<td>RAD</td>
<td>Rapid Application Development</td>
</tr>
<tr>
<td>DNS</td>
<td>Domain Name System</td>
</tr>
<tr>
<td>AS</td>
<td>Autonomous System</td>
</tr>
<tr>
<td>ISP</td>
<td>Internet Service Provider</td>
</tr>
<tr>
<td>DB</td>
<td>Database</td>
</tr>
<tr>
<td>k-NN</td>
<td>K Nearest Neighbor</td>
</tr>
<tr>
<td>TP</td>
<td>True Positive</td>
</tr>
<tr>
<td>FP</td>
<td>False Positive</td>
</tr>
<tr>
<td>MCC</td>
<td>Matthew Correlation Coefficient</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Characteristic</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

1.1 Introduction

People nowadays are fully dependent towards the Internet. This is because by using the Internet, everyone will be able to access everything online anywhere and anytime. People communicate with each other online, do online transaction, store every type of data online instead of doing everything offline which will need huge storage size and there will be high possibility that the data will either go missing or corrupt.

Although there are so many advantages of doing everything online, not all online activities are safe. Based on the statistical studies made by some web sites, there have been fluctuating results throughout year 2017 on the affected users from malicious attacks (Roman Unuchek, Fedor Sinitsyn, Denis Parinov, 2017). Whereas, for the type of new threats are also increasing, for example ransomware and phishing attacks (Gammons, 2017). There are also reports of the results from the comparison for year 2015, 2016 and 2017, where the top attack techniques were recorded. The most popular technique for year 2015 is unknown attack and followed by SQL injection. In year 2016, the most popular technique is also unknown attack followed by account hacking. In year 2017, the most popular technique is the malware attack (Passeri, 2018).

With the growing numbers of malicious threats, many systems have been developed to detect any cybercrimes and eventually to get rid of them. However, the numbers of sophisticated hackers are also growing and they always make sure to be able to attack anything they want as long as they get what they need using every possible method. Hence, this project is to develop a system for detecting any kind of malicious websites as it is one of the easiest and most common method for cyber criminals to attack every computer user.
Malicious websites are one of the ways for the computers to get infected with. This easily happens when the attacker links a user to a website that looks exactly like the familiar sites for instance Google, PayPal or Gmail when actually they are the scammer’s site. Users will often enter their username or password on the malicious site and as the result, the attacker will then have the complete control over the users’ account.

Most people are unaware of the fact that they do not have to intentionally download a malicious attachment in order to compromise the computer’s security. Malicious websites and drive-by downloads are just the two ways that the security can become compromised by doing nothing more than visiting a website. A malicious website will attempt to install malware onto the devices without users being aware of it or asking for permission first to either disrupt computer operation or gathering personal information or in a worst-case scenario, to gain total access to the machine.

1.2 Problem Statement

There are two problems identified which leading to the development of this project. The first problem is that the website security is not guaranteed (Compromise, D., State, T. H. E., & Security, O. F. (2018). Trustwave Global Security Report Introduction the State of Security). All websites are not always secured and even if they have put the necessary security measures, they are still easily hacked by hackers for various kind of purposes.

The next problem is the exploitation of users’ important credentials (Compromise, D., State, T. H. E., & Security, O. F. (2018). Trustwave Global Security Report Introduction the State of Security). It can be an instant exploitation or the attacker will use the information for future attack. This will not only be a security breach but also a few more issues will be stemmed from that much of information.
1.3 Objectives

The objectives are important to achieve the project development goals. The objectives of this project are:

i. To identify the features for malicious websites.
ii. To evaluate the proposed system.

1.4 Scope

For this project, there are four scopes to be covered throughout the research project development.

The first scope is the three techniques that will be discussed further in the research which are signature-based detection technique, heuristic-based detection technique and N-gram technique.

The second scope is the five features to be used for the result evaluation which are URL feature, host-based feature, content-based feature, graph-based feature and blacklist feature.

The third scope is the five algorithms to be compared in this research which are random forests, naïve bayes, k-nearest neighbors, support vector machine and adaptive boost.

The last scope is the 100 of the datasets to be used in this research which consisting of 50 malicious URLs and 50 benign URLs.

1.5 Significance

The malicious website detection tool will be very useful to all computer users since everyone wants their confidential to be protected and to be ensured that the security of their systems and the websites that they are searching for are strong so they do not have to be anxious about using the computers and surfing through the Internet. They do not have to worry anymore about getting their privacy and confidential being violated.
REFERENCES


Heiderich, M., Frosch, T., & Holz, T. (2011). IceShield: Detection and mitigation of malicious websites with a frozen DOM. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 6961 LNCS, pp. 281–300). https://doi.org/10.1007/978-3-642-23644-0_15


