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)

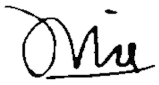
Full Name : EN. MOHD FAIZAL AB RAZAK

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

Laman web berniat jahat adalah antara ancaman keselamatan utama di Internet. Ancaman ini telah wujud selama bertahun-tahun namun penyelesaian terbaik untuk mengatasinya tidak diamalkan oleh orang ramai. Kebanyakan kaedah sedia ada untuk mengesan laman web berniat jahat lebih menumpukan perhatian terhadap serangan tertentu. Walau bagaimanapun, serangan semakin kompleks dan penggodam menjadi lebih bijak dengan teknik campuran mereka untuk mengelakkan daripada dikesan. Dalam tesis ini, satu kaedah akan diperkenalkan. Dengan kaedah yang sedia ada dipertimbangkan, kaedah yang digunakan untuk projek ini adalah dengan menggunakan pengesanan berasaskan heuristik dengan teknik pembelajaran mesin atau lebih dikenali sebagai 'machine learning' dan ciri yang akan digunakan bersama teknik ini adalah ciri berdasarkan URL. Tujuan kaedah ini adalah mengklasifikasikan laman web yang normal dan berniat jahat dengan menggunakan mesin pembelajaran dan kemudian secara automatik akan mengesan laman web yang berniat jahat. Dengan menggunakan kaedah ini juga dapat memastikan ketepatan pengesanan yang tinggi dan semua laman web yang berniat jahat boleh dikesan walaupun baru dikemaskini oleh penggodam. Kesimpulannya, kaedah yang dicadangkan merupakan salah satu cara yang paling berkesan untuk mengesan laman web berniat jahat dan mudah untuk dilaksanakan.

ABSTRACT

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

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

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

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

Breiman, L. E. O. (2001). Random Forest（LeoBreiman）.pdf, 5–32. https://doi.org/10.1023/A:1010933404324

Chen, B., & Chen, M. (2011). 65-G00007.pdf, *1*, 301–305.

Choi, H., Zhu, B. B., & Lee, H. (2011). Detecting malicious web links and identifying their attack types. *WebApps*, 11. https://doi.org/10.1109/IUCS.2010.5666254

Cho, M. Y., & Hoang, T. T. (2017). Feature Selection and Parameters Optimization of SVM Using Particle Swarm Optimization for Fault Classification in Power Distribution Systems. *Computational Intelligence and Neuroscience*, *2017*. https://doi.org/10.1155/2017/4135465

Chong, C., Liu, D., & Lee, W. (2009). Malicious url detection, 1–4. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.278.4330&rep=rep1&type=pdf

Collins, M., & Singer, Y. (1999). Unsupervised models for named entity classification. *Proceedings of EMNLP/VLC-99*, 100–110. https://doi.org/10.1.1.114.3629

Connor, M. F. O., Moosaei, M., Bixler, R., & Iqbal, T. (n.d.). Systems Analysis of the WEKA Machine Learning Workbench for Affective Computing.

Cova, M., Kruegel, C., & Vigna, G. (2010). Detection and analysis of drive-by-download attacks and malicious JavaScript code. In *Proceedings of the 19th international conference on World wide web - WWW '10* (p. 281). https://doi.org/10.1145/1772690.1772720

Dietterich, T. G. (1990). Dietterich00.pdf. Retrieved from http://www.eecs.wsu.edu/~holder/courses/CptS570/fall07/papers/Dietterich00.pdf

Eshete, B., Villafiorita, A., & Weldemariam, K. (2013). BINSPECT: Holistic analysis and detection of malicious web pages. In *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering* (Vol. 106 LNICS, pp. 149–166). https://doi.org/10.1007/978-3-642-36883-7_1

Gammons, B. (2017). 6 Must-Know Cybersecurity Statistics for 2017 | Barkly Blog. Retrieved from https://blog.barkly.com/cyber-security-statistics-2017

Genuer, R., Poggi, J., & Tuleau-malot, C. (2012). Variable selection using Random Forests To cite this version : *Pattern Recognition Letters*, *31*(14), 2225–2236.

Guan, C., Qin, S., Ling, W., & Ding, G. (2016). Apparel recommendation system evolution: an empirical review. *International Journal of Clothing Science and Technology*, *28*(6), 854–879. https://doi.org/10.1108/IJCST-09-2015-0100

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

Hou, Y. T., Chang, Y., Chen, T., Laih, C. S., & Chen, C. M. (2010). Malicious web content detection by machine learning. *Expert Systems with Applications*, *37*(1), 55–60. https://doi.org/10.1016/j.eswa.2009.05.023

Joachims, T. (2002). Learning to Classify Text Using Support Vector Machines, (Vapnik 1995). https://doi.org/10.1007/978-1-4615-0907-3

Kazemian, H. B., & Ahmed, S. (2015). Comparisons of machine learning techniques for detecting malicious webpages. *Expert Systems with Applications*, *42*(3), 1166–1177. https://doi.org/10.1016/j.eswa.2014.08.046

Keller, J. M., & Gray, M. R. (1985). A Fuzzy K-Nearest Neighbor Algorithm. *IEEE Transactions on Systems, Man and Cybernetics*, *SMC-15*(4), 580–585. https://doi.org/10.1109/TSMC.1985.6313426

Klusowski, J. M. (2018). Complete Analysis of a Random Forest Model, *13*, 1063–1095. https://doi.org/arXiv:1805.02587v2

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review*, *26*(3), 159–190. https://doi.org/10.1007/s10462-007-9052-3

Lee, C., & Lee, G. G. (2006). Information gain and divergence-based feature selection for machine learning-based text categorization. *Information Processing and Management*, *42*(1 SPEC. ISS), 155–165. https://doi.org/10.1016/j.ipm.2004.08.006

Lin, S. W., Ying, K. C., Chen, S. C., & Lee, Z. J. (2008). Particle swarm optimization for parameter determination and feature selection of support vector machines. *Expert Systems with Applications*, *35*(4), 1817–1824. https://doi.org/10.1016/j.eswa.2007.08.088

Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2011). Learning to detect malicious URLs. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 1–24. https://doi.org/10.1145/1961189.1961202

Manevitz, L. M., & Yousef, M. (2001). One-Class SVMs for Document Classification. *Journal of Machine Learning Research 2*, *2*, 139–154. https://doi.org/10.1162/15324430260185574

Mittal, S. (2016). Machine Learning Approach for Classifying Malicious URLs, *4*(4), 622–629.

Passeri, P. (2018). 2017 Cyber Attacks Statistics. Retrieved from https://www.hackmageddon.com/2018/01/17/2017-cyber-attacks-statistics/

Patil, D. R., & Patil, J. B. (2018). Malicious URLs detection using decision tree classifiers and majority voting technique. *Cybernetics and Information Technologies*, *18*(1), 11–29. https://doi.org/10.2478/cait-2018-0002

Pavlidis, P., Wapinski, I., & Noble, W. S. (2004). Support vector machine classification on the web. *Bioinformatics*, *20*(4), 586–587. https://doi.org/10.1093/bioinformatics/btg461

Roman Unuchek, Fedor Sinitsyn, Denis Parinov, A. L. (2017). IT threat evolution Q3 2017. Statistics. Retrieved from https://securelist.com/it-threat-evolution-q3-2017-statistics/83131/

Rouse, Margaret, and Ed Burns. "What Is Machine Learning (ML)? - Definition from WhatIs.com." SearchEnterpriseAI, May 2018, searchenterpriseai.techtarget.com/definition/machine-learning-ML.

Sahoo, D., Liu, C., & Hoi, S.C. (2017). Malicious URL Detection using Machine Learning: A Survey. *CoRR, abs/1701.07179*.

Sayamber, A., & Dixit, A. (2014). Malicious URL Detection and Identification. *Journal of Computer Applications (0975–8887)*, *99*(17), 17–23. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.678.5517&rep=rep1&type=pdf

Symantec. (2016). Internet Security Threat Report Symantec 2016. *Network Security*, *2016*, 81. Retrieved from https://www.symantec.com/content/dam/symantec/docs/reports/istr-21-2016-en.pdf

Vanhoenshoven, F., Nápoles, G., & Falcon, R. (2016). Detecting malicious URLs using machine learning techniques. *(Ssci), 2016 Ieee …*, (December). Retrieved from http://ieeexplore.ieee.org/abstract/document/7850079/

Witten, H. I., Frank, E., Trigg, L., Hall, M., Holmes, G., & Cunningham, S. J. (1999). Uow-Cs-Wp-1999-11.Pdf.

Xu, S., Bylander, T., Maynard, H. B., Sandhu, R., & Xu, M. (2014). Detecting and Characterizing Malicious Websites.

Yashaswini, J., Varshney, G., & Nagaraju, M. (2018). Comparative Analysis of Algorithms Detecting Malicious URL ' s ., *3*(3), 302–305.

Z. (2014). *U.S. Patent No. US8850570B1*. Washington, DC: U.S. Patent and Trademark Office.