ANTI-PHISHING MODEL FOR PHISHING WEBSITES DETECTION: USING PRUNING DECISION TREE

AHMED I. M. ABUNADI

A project report submitted in partial fulfilment of the
requirements for the award of the degree of
Master of Computer Science (Information Security)

Faculty Computing
Universiti Teknologi Malaysia

JUNE 2013

I dedicate this project to my beloved parents and my beloved wife , thank you for the moral support you have given me throughout my academic life.

To my respected supervisor, Dr. Anazida Zainal

To my beloved country, Palestine

To all my brothers and sisters

To all my friends

ACKNOWLEDMENT

First and foremost, all praise and thanks are due to Allah, and peace and blessings be upon his Messenger, Mohammed (Peace Be Upon Him). Next, I would like to express heartfelt gratitude to my supervisor **Dr. Anazida Zainal** for her constant support during my study at UTM. She inspired me greatly to work in this project. Her willingness to motivate me contributed tremendously to our project. I have learned a lot from her and I am fortunate to have her as my mentor and supervisor. Special thanks to my parents who cheered me on from the beginning of my study. Thanks very much for my dear wife who helped, encouraged me and provide a suitable environment for study. Last but not least, I am grateful to my beloved family and all my friends for their warm encouragements and supports.

Besides, I would like to thank the authority of Universiti Teknologi Malaysia (UTM) for providing me with a good environment and facilities such as Computer laboratory to complete this project with software which I need during process.

ABSTRAK

Sebagai satu bentuk baru perisian berniat jahat, laman web phishing sering muncul pada tahun-tahun kebelakangan ini, yang menyebabkan kemudaratan yang besar kepada perkhidmatan kewangan dalam talian dan keselamatan data. Banyak kajian telah dilakukan untuk mengurangkan laman web phishing. Kebanyakan kajian menggunakan laman web yang mempunyai ciri-ciri web phishing untuk proses penyiasatan. Banyak kajian telah diakukan untuk mengurangkan laman web phishing. Kebanyakan kajian menggunakan laman web yang bercirikan web phishing bagi proses penyiasatan. Sesetengah ciri-ciri ini tidak mempunyai nilai yang ketara kepada nisbah ketepatan yang boleh menjejaskan prestasi dari segi masa pengiraan. Di samping itu, ciri-ciri ini boleh didapati di kedua-dua phishing dan bukan phishing dengan nilai-nilai yang sama yang akan menjejaskan ketepatan pengesanan. Dalam kajian ini, ciri-ciri laman web phishing telah dibincangkan secara terperinci. Di samping itu, kajian ini menunjukkan ciri-ciri baru untuk mengesan laman web phishing menggunakan kaedah Pemangkasan Keputusan Pokok untuk mengimbangi masa pengiraan dan nisbah ketepatan. Pemangkasan Ralat Pesimis digunakan sebagai algoritma cantas untuk mencantas dedaun pokok keputusan tanpa menjejaskan ketepatan. Pengkategorian laman web Phising adalah satu lagi objektif bagi kajian ini, dan bertujuan untuk memberi tips khusus dan penting untuk meningkatkan tahap kesedaran di kalangan pengguna bagi setiap kategori.

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

PROJECT OVERVIEW

1.1 Introduction

Social engineering attacks targeting users not computers or systems and it is designed to elicit sensitive or confidential information from users. Most of social engineering attacks are classified as phishing attacks. And there are different techniques for phishing in order to deceive the users such as phishing by email, instant messages, SMS and website. These techniques help the phisher to trick the users for various purposes such illegal profit, get personal data or sensitive information.

In general, we can define phishing as an internet crime whereby the attackers use social engineering in order to fraud users. And this fraud can be done by sending emails, through advertisements on websites or even by phone calls to get the attention of users. Simply, the goal is to allure users to phishing websites that mimics a legitimate websites to ruse users in order to get their sensitive information such as passwords, credits card, e-bank account, etc. As a result the attacker can use their information to do what he wants to do such as illegal profit or impersonate them (Liu et al., 2010).

Recently, phishing websites becomes more sophisticated which is difficult to judge if it is a legitimate website or not. According to the Anti-Phishing Working Group (2012), 53,939 unique phishing websites were detected in the first half of 2012. These websites can allure a lot of users and trick them in sake of taking their sensitive information especially normal users who do not look on details when surfing websites. So it is obvious that we need efficient solutions to mitigate the danger of phishing websites and aware the users more about this threat in order to protect themselves in future.

1.2 Problem Background

Phishing attacks started with a spoofed email masquerading as authorize sender. These emails impersonate credit card companies, ecommerce websites, IT service provider or brand names of banks. And these emails have a text or images which are trying to convince the user by two ways either by generating trepidation or generating exhilaration in order to a lure the user to click on any links inside that email which it will redirect him to a phishing website. Of course both the phishing email and Web site are replica copy of the original website to convince the user to disclose his sensitive information.

There are a lot of researches (Afroz and Greenstadt, 2011, Alnajim and Munro, 2009, Lakshmi and Vijaya, 2012, Liu et al., 2010, Maher Aburrous, 2010, Weiwei et al., 2012) have been conduct in order to detect or prevent or mitigate phishing websites and they proposed different solution approaches. First solution approach is based on blacklist approach which is very simple and inefficient because it depends on a remote database to check whether the website is a phishing or a legitimate website. The second approach is more efficient because it is based on intelligent solutions. So it will extract some features from websites and pass the data to an algorithm to detect phishing websites. Finally, the third approach combines the black list and intelligent approaches. This approach first checks the black list, and if

the URL is not in the list, it will execute the intelligent algorithm for detection. Later, the list will be updated. Therefore, the computation time can be reduced.

Intelligent solutions use different rules to detect phishing websites and different algorithms to train the reference model or the solution. And some of these solutions are targeting only one type of phishing websites such as e-banking websites. In addition, most of these solutions use a lot of rules in order to detect phishing websites but unfortunately some rules have less impact on result. Therefore, it may give high percentage of false positive.

1.3 Problem Statement

Most of current anti-phishing solutions do not include categorization of phishing websites. Furthermore these solutions use many rules which increase the time computation and the complexity of the classifier. In addition, some of the rules might increase the percentage of false positive.

1.4 Purpose of Study

This research is going to focus on identifying most common features of phishing websites in order to enhance detection accuracy, decrease false positive and categorize the phishing websites into four categories (e-banking, online shopping, IT services and others).

1.5 Project Objectives

The objectives of this study are listed below:

I. Find out new features for phishing websites in order to enhance detection function.
II. Reduce the computation time and false positive using Pruning Decision Tree in order to make the solution more efficient.
III. Categorize phishing websites according to extracted features to four categories (e-banking, online shopping, IT service provider and others).

The purpose of using Pruning Decision Tree is to help us to discard any rule which does not have effect on the result or less impact. Using Pruning Decision Tree will decrease the difficulty of the final classifier and enhance the prediction accuracy and as a result it will reduce the size of the tree.

1.6 Scope of Study

The scopes of this project that specify the boundaries are listed below:

I. The study focuses on developing an intelligent model for phishing websites detection using Pruning Decision Tree.
II. Classifying the phishing websites according to some features inside HTML code.
III. The study will use some dataset from online resources in order to train the model (http://www.phishtank.com , http://www.antiphishing.org).

1.7 Significant of Research Study

The purpose of this study is to develop more efficient model for phishing websites detection using Pruning Decision Tree to reduce false positive. Another purpose is categorize phishing websites in order to increase the awareness level among usersthrough giving them tips based on the type of phishing website.

1.8 Organization of Report

The thesis consists of 4 chapters. Chapter one describes the introduction, problem background, problem statements, objectives, the scope of study and significant of study. Chapter twois presenting literature review. Chapter three includes the project methodology. Chapter four presents the findings of the initial result.

REFERENCES

ABURROUS, M., HOSSAIN, M. A., THABATAH, F. & DAHAL, K. Intelligent phishing website detection system using fuzzy techniques. Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on, 2008. IEEE, 1-6.

AFROZ, S. & GREENSTADT, R. PhishZoo: Detecting Phishing Websites by Looking at Them. Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on, 18-21 Sept. 2011 2011. 368-375.

BLYTHE, M., PETRIE, H. & CLARK, J. A. F for fake: Four studies on how we fall for phish. PART 5--------Proceedings of the 2011 annual conference on Human factors in computing systems, 2011. ACM, 3469-3478.

CHEN, K. T., CHEN, J. Y., HUANG, C. R. & CHEN, C. S. 2009. Fighting phishing with discriminative keypoint features. *Internet Computing, IEEE*, 13**,** 56-63.

CYBERSOURCE. 2012. *Online Payment Fraud Trend Report* [Online]. CyberSource. [Accessed 5/11/2012 2012].

DEY, S. K., NABI, M. N. & ANWER, M. Challenges in building trust in B2C e-Commerce and proposal to mitigate them: developing countries perspective. Computers and Information Technology, 2009. ICCIT'09. 12th International Conference on, 2009. IEEE, 581-586.

DHAMIJA, R., TYGAR, J. D. & HEARST, M. Why phishing works. Proceedings of the SIGCHI conference on Human Factors in computing systems, 2006. ACM, 581-590.

DRAKE, C. E., OLIVER, J. J. & KOONTZ, E. J. Anatomy of a phishing email. Conference on Email and Anti-Spam, 2004.

DUNLOP, M., GROAT, S. & SHELLY, D. GoldPhish: Using Images for Content-Based Phishing Analysis. Internet Monitoring and Protection (ICIMP), 2010 Fifth International Conference on, 2010. IEEE, 123-128.

ESPOSITO, F., MALERBA, D., SEMERARO, G. & KAY, J. 1997. A comparative analysis of methods for pruning decision trees. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 19**,** 476-491.

FU, A. Y., WENYIN, L. & DENG, X. 2006. Detecting phishing web pages with visual similarity assessment based on earth mover's distance (EMD). *Dependable and Secure Computing, IEEE Transactions on,* 3**,** 301-311.

GANG LIU, B. Q., LIU WENYIN 2010. Automatic Detection of Phishing Target from Phishing Webpage. *2010 International Conference on Pattern Recognition.* Istanbul.

JULIE S. DOWNS, M. B. H., LORRIE FAITH CRANOR 2006. Decision Strategies and Susceptibility to Phishing. *Proc. the 2nd symposium on usable privacy and security*. New York, USA: ACM Press.

MAHER ABURROUS, M. A. H., KESHAV DAHAL, FADI THABTAH 2010. Predicting Phishing Websites using Classification Mining Techniques with Experimental Case Studies. *Seventh International Conference on Information Technology.*

MINGERS, J. 1989. An empirical comparison of pruning methods for decision tree induction. *Machine Learning,* 4**,** 227-243.

MOUSTAKAS, E., DUQUENOY, P. & RANGANATHAN, C. Phish or Treat? Understanding the Tactics and Responses to Electronic Identity Theft on the Internet. Proceedings of the 5th European Conference on i-Warfare and Security. Academic Conferences Limited, 239.

OLSHEN, L. B. J. H. F. R. A. & STONE, C. J. 1984. Classification and Regression Trees. *Wadsworth International Group*.

PAN, Y. & DING, X. Anomaly based web phishing page detection. Computer Security Applications Conference, 2006. ACSAC'06. 22nd Annual, 2006. IEEE, 381-392.

PHILIPPSOHN, S. 2001. Trends In Cybercrime — An Overview Of Current Financial Crimes On The Internet. *Computers & Security,* 20**,** 53-69.

PUNDIR, P. & GOMANSE, V. 2011. ATTACK VECTORS USED IN FRAUDULENCE CONNECTION DURING ONLINE TRANSACTIONS. *International Journal,* 3.

QI, M. & YANG, C. Research and Design of Phishing Alarm System at Client Terminal. Services Computing, 2006. APSCC'06. IEEE Asia-Pacific Conference on, 2006. IEEE, 597-600.

QUINLAN, J. R. 1987. Simplifying decision trees. *International journal of man-machine studies,* 27**,** 221-234.

QUINLAN, J. R. 1993. *C4. 5: programs for machine learning*, Morgan kaufmann.

ROKACH, L. & MAIMON, O. 2005. Decision trees. *Data Mining and Knowledge Discovery Handbook***,** 165-192.

WATTERS, P. A. Why do users trust the wrong messages? A behavioural model of phishing. eCrime Researchers Summit, 2009. eCRIME '09., Sept. 20 2009-Oct. 21 2009 2009. 1-7.

WEIWEI, Z., QINGSHAN, J. & TENGKE, X. An Intelligent Anti-phishing Strategy Model for Phishing Website Detection. Distributed Computing Systems Workshops (ICDCSW), 2012 32nd International Conference on, 18-21 June 2012 2012. 51-56.

ZDZIARSKI, J., YANG, W. & JUDGE, P. Approaches to Phishing Identification using Match and Probabilistic Digital Fingerprinting Techniques. Proceedings of the MIT Spam Conference, 2006. 1115-1122.

AL SHALABI, L. & SHAABAN, Z. Normalization as a preprocessing engine for data mining and the approach of preference matrix. Dependability of Computer Systems, 2006. DepCos-RELCOMEX'06. International Conference on, 2006. IEEE, 207-214.

HAN, J., KAMBER, M. & PEI, J. 2006. *Data mining: concepts and techniques*, Morgan kaufmann.

RAPIDMINER, R. 2009. Open Source Data Mining.

SULLIVAN, D. 2002. How search engines work. *Search Engine Watch,* 14.

BRAMER, M. 2007. *Principles of data mining*, Springer.

KANG, J. & LEE, D. Advanced white list approach for preventing access to phishing sites. Convergence Information Technology, 2007. International Conference on, 2007. IEEE, 491-496.

KIM, Y.-G., CHO, S., LEE, J.-S., LEE, M.-S., KIM, I. H. & KIM, S. H. 2008. Method for evaluating the security risk of a website against phishing attacks. *Intelligence and Security Informatics.* Springer.

MORENO, H. 2012. *Machine Learning for Student Modeling.* WORCESTER POLYTECHNIC INSTITUTE.

QUINLAN, J. R. 1993. *C4. 5: programs for machine learning*, Morgan kaufmann.

ROKACH, L. & MAIMON, O. 2005. Top-down induction of decision trees classifiers-a survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on,* 35**,** 476-487.

RUGGIERI, S. Subtree Replacement in Decision Tree Simplification.

SANGLERDSINLAPACHAI, N. & RUNGSAWANG, A. Using domain top-page similarity feature in machine learning-based web phishing detection. Knowledge Discovery and Data Mining, 2010. WKDD'10. Third International Conference on, 2010. IEEE, 187-190.

USHARANI, C. & CHANDRASEKARAN, R. 2010. Course planning of higher education to meet market demand by using data mining techniques–a case of a Technical University in India. *International Journal of Computer Theory and Engineering,* 2**,** 1793-8201.

WEIWEI, Z., QINGSHAN, J. & TENGKE, X. An Intelligent Anti-phishing Strategy Model for Phishing Website Detection. Distributed Computing Systems Workshops (ICDCSW), 2012 32nd International Conference on, 18-21 June 2012 2012. 51-56.

WHITE, A. P. & LIU, W. Z. 1994. Technical note: Bias in information-based measures in decision tree induction. *Machine Learning,* 15**,** 321-329.

ZHANG, H., LIU, G., CHOW, T. W. & LIU, W. 2011. Textual and visual content-based anti-phishing: a Bayesian approach. *Neural Networks, IEEE Transactions on,* 22**,** 1532-1546.