Faculty of Computer Science and Information Technology

Phishing Detection with Identity Keywords and Target Domain Name

Colin Tan Choon Lin

Master of Science
2015

Phishing Detection with Identity Keywords and Target Domain Name

Colin Tan Choon Lin

A thesis submitted

In fulfillment of the requirements for the degree of

Master of Science

Faculty of Computer Science and Information Technology

UNIVERSITI MALAYSIA SARAWAK

2015

UNIVERSITI MALAYSIA SARAWAK

Grade: _____

Please tick (√)

Final Year Project Report	☐
Masters	☑
PhD	☐

DECLARATION OF ORIGINAL WORK

This declaration is made on the ……..day of……..2015.

Student's Declaration:

I, COLIN TAN CHOON LIN, 14020073, FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY hereby declare that the work entitled, PHISHING DETECTION WITH IDENTITY KEYWORDS AND TARGET DOMAIN NAME is my original work. I have not copied from any other students' work or from any other sources except where due reference or acknowledgement is made explicitly in the text, nor has any part been written for me by another person.

_____ _____

Date submitted Colin Tan Choon Lin (14020073)

Supervisor's Declaration:

I, DR. CHIEW KANG LENG hereby certifies that the work entitled, PHISHING DETECTION WITH IDENTITY KEYWORDS AND TARGET DOMAIN NAME was prepared by the above named student, and was submitted to the "FACULTY" as a partial fulfillment for the conferment of MASTER OF SCIENCE IN COMPUTER SCIENCE, and the aforementioned work, to the best of my knowledge, is the said student's work.

Received for examination by: _____ Date: _____

(Dr. Chiew Kang Leng)

I declare this Project/Thesis is classified as (Please tick (√)):

☐ **CONFIDENTIAL** (Contains confidential information under the Official Secret Act 1972)*****
☐ **RESTRICTED** (Contains restricted information as specified by the organisation where research was done)*****
☑ **OPEN ACCESS**

Validation of Project/Thesis

I therefore duly affirm with free consent and willingly declare that this said Project/Thesis shall be placed officially in the Centre for Academic Information Services with the abiding interest and rights as follows:

- This Project/Thesis is the sole legal property of Universiti Malaysia Sarawak (UNIMAS).
- The Centre for Academic Information Services has the lawful right to make copies for the purpose of academic and research only and not for other purpose.
- The Centre for Academic Information Services has the lawful right to digitalise the content for the Local Content Database.
- The Centre for Academic Information Services has the lawful right to make copies of the Project/Thesis for academic exchange between Higher Learning Institute.
- No dispute or any claim shall arise from the student itself neither third party on this Project/Thesis once it becomes the sole property of UNIMAS.
- This Project/Thesis or any material, data and information related to it shall not be distributed, published or disclosed to any party by the student except with UNIMAS permission.

Student's signature: _____ (Date:)

Supervisor's signature: _____ (Date:)

Current Address:
687, Heights Drive, Jalan Stutong, Lorong Stutong 11, 93350 Kuching, Sarawak, Malaysia.

Notes: ***** If the Project/Thesis is **CONFIDENTIAL** or **RESTRICTED**, please attach together as annexure a letter from the organisation with the period and reasons of confidentiality and restriction.

[The instrument is duly prepared by The Centre for Academic Information Services]

LIST OF PUBLICATIONS

1. Tan, C. L., Chiew, K. L., & Sze, S. N. (2014). Phishing website detection using URL-assisted brand name weighting system. In *2014 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Kuching, Malaysia, 1-4 December* (pp. 54–59). http://doi.org/10.1109/ISPACS.2014.7024424

2. Tan, C. L., Chiew, K. L., & Sze, S. N. (forthcoming 2016). Phishing Webpage Detection Using Weighted URL Tokens for Identity Keywords Retrieval. In *9th International Conference on Robotics, Vision, Signal Processing & Power Applications (RoViSP 2016), Penang, Malaysia, 2-3 February*.

ACKNOWLEDGEMENT

Throughout the course of my Master's research, several individuals have been a great source of help and encouragement to me. First of all, I would like to take this opportunity to express my sincere gratitude and appreciation to my supervisor Chiew Kang Leng for his patience and immense knowledge in the art of research. His constant assurance in the midst of my confusion and disappointment has always inspired me to persevere on and strive harder.

My sincere gratitude also goes towards my beloved parents, Tan Tze Yong and Kuek Eng Lan for their unceasing love and cares throughout my life. Their constant support and prayers for me has blessed me tremendously.

I also wish to extend my appreciation to Lauretha Rura, for patiently bearing with me in life and supporting me all the way as I pursue my Master's degree. It is truly a great joy to journey together with her through the ups and downs of my life.

Not forgetting my fellow research mate, Chang Ee Hung who has helped me greatly during the initial stages of dataset collection. His companionship and willingness to share his knowledge has contributed much to my research progress.

In addition, I also wish to thank the Ministry of Higher Education, Malaysia for assisting me financially through the MyBrain15 scholarship and the Fundamental Research Grant Scheme 2/2013 [Grant No: FRGS/ICT07(01)/1057/2013(03)].

Finally, I would like to express my highest gratitude to my God, Jesus Christ. It is by His favour and goodness that brings me through all the challenges and difficult moments in my life.

ABSTRACT

This thesis describes the research work carried out to address the problem of phishing detection and the weaknesses in existing anti-phishing methods. Phishing works by luring users to counterfeit websites, where highly confidential credentials are requested. To safeguard Internet users against phishing attacks, a hybrid anti-phishing method consisting of text-based, search engine-based and identity-based methods are proposed, where the differences between the target and actual identities of a webpage are exploited for classification. The proposed method can be divided into three phases. The first phase extracts identity keywords from the textual contents of the website, where a novel weighted URL tokens system based on the N-gram model is proposed. The second phase finds the target domain name by using a search engine, and the target domain name is selected based on identity-relevant features. In the final phase, a 3-tier identity matching system exploits indirect identity relationships to conclude the legitimacy of the query webpage. Experiments were conducted over 10,000 datasets, where true positive rate of 99.68% and true negative rate of 92.52% were achieved. Benchmarking results also suggest that the proposed method achieves comparable overall accuracy with three selected conventional methods. In summary, the proposed method has the key advantage of identifying phishing webpages accurately. This key advantage is highly desirable in anti-phishing applications.

Pengesanan Laman Web Palsu Dengan Kata Kunci dan Nama Domain Sasaran

ABSTRAK

Tesis ini menerangkan kajian yang telah dijalankan untuk menangani masalah pengesanan laman web palsu dan kelemahan-kelemahan dalam sistem yang sedia ada. Jenayah siber ini dijalankan dengan mengumpan pengguna Internet ke laman web palsu untuk mencuri maklumat rahsia pengguna. Untuk melindungi pengguna Internet, satu kaedah hibrid yang terdiri daripada kaedah teks, enjin carian dan identiti telah diperkenalkan. Kaedah ini mengeksploitasi perbezaan antara identiti sasaran dengan identiti sebenar untuk mengesahkan ketulenan sesuatu laman web. Kaedah tersebut dibahagikan kepada tiga fasa. Fasa pertama mengekstrak kata kunci daripada kandungan teks laman web dengan menggunakan satu sistem pemarkahan token URL berasaskan model N-gram. Fasa kedua mencari nama domain sasaran dengan menggunakan enjin carian, diikuti dengan pemilihan nama domain sasaran berdasarkan ciri-ciri identiti. Dalam fasa terakhir, satu sistem perbandingan identiti secara berperingkat akan mengeksploitasi hubungan identiti secara tidak langsung untuk mengesahkan ketulenan laman web yang diuji. Eksperimen yang dijalankan ke atas 10,000 laman web telah berjaya mengklasifikasikan 99.68% laman web palsu dan 92.52% laman web tulen. Keputusan eksperimen juga menunjukkan kaedah yang kami perkenalkan mencapai ketepatan keseluruhan yang setanding dengan tiga sistem lain yang sedia ada. Kesimpulannya, kaedah yang kami perkenalkan mempunyai kelebihan dalam mengenalpasti laman web palsu dengan tepat. Kelebihan ini sangat diperlukan dalam aplikasi-aplikasi pengesanan laman web palsu.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACC	Accuracy
API	Application Programming Interface
APWG	Anti-Phishing Working Group
ASCII	American Standard Code for Information Interchange
CANTINA	Carnegie Mellon Anti-phishing and Network Analysis Tool
CBD	Content-Based Detection
ccTLD	Country Code Top Level Domain
CPU	Central Processing Unit
CSS	Cascading Style Sheets
DNS	Domain Name System
DOM	Document Object Model
EMD	Earth Mover's Distance
FN	False Negatives
FNR	False Negative Rate
FP	False Positives
FPR	False Positive Rate
gTLD	Generic Top Level Domain
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secure

IANA	Internet Assigned Numbers Authority
IP	Internet Protocol
KNN	K-Nearest Neighbour
LHS	Left Hand Side
MCC	Matthews Correlation Coefficient
MCDM	Multi-Criteria Decision Making
MLAPT	Machine Learning Anti-Phishing Technique
MW	Megawatt
NER	Named Entity Recognizer
OCR	Optical Character Recognition
POS	Part-Of-Speech
RAM	Random Access Memory
RHS	Right Hand Side
SEO	Search Engine Optimization
SIFT	Scale Invariant Feature Transform
SLD	Second Level Domain
SLN	Semantic Link Network
SMTP	Simple Mail Transfer Protocol
SPF	Sender Policy Framework
SSL	Secure Sockets Layer
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
TLD	Top Level Domain
TN	True Negatives

TNR	True Negative Rate
TP	True Positives
TPR	True Positive Rate
URL	Uniform Resource Locator
XPath	XML Path Language

CHAPTER 1

INTRODUCTION

1.0 Research Background

In this modern age of information technology, consumers are dealing with more products and services through the online channel. Most transactions can now be performed through the online gateway, which require users to key in some form of authorisation credentials. Therefore, having multiple online accounts (e.g., email account, banking account, social networking account, etc.) has become a norm for most people. This technological trend is exposing many Internet users to a rising threat of online identity theft known as phishing.

Phishing is a social engineering scheme launched by profit-driven criminals to coerce unsuspecting users into disclosing their confidential information on counterfeit websites. Phishers usually entice victims to the phishing website by sending emails containing the fraudulent URL and some threatening messages (e.g., account termination, data loss, etc.) to instil a sense of urgency. At the phishing website, the phishers will capture information submitted by the victims and use it to gain access to the victims' actual accounts and monetary resources.

The threat of phishing has not diminished even after a decade, as phishers continue to exploit the human factors. Based on the human behaviour studies by Dhamija et al. (2006) and Alsharnouby et al. (2015), the success rate of phishing attack on a typical Internet user is fairly high. Both studies suggest that most users do not truly understand the meaning of important security indicators such as the Secure Sockets Layer (SSL) protocol and digital

certificate on the browser address bar. Some users might also be confused on how a legitimate URL is supposed to resemble, thus they rely on the webpage contents to determine its genuineness (Mohammad, Thabtah, & McCluskey, 2015).

The severity of phishing threats in recent years has grown considerably, based on a few statistics gathered from security organizations. The Anti-Phishing Working Group (APWG) has observed a total of 42,212 unique phishing websites in June 2014, as compared to 38,110 in June 2013 (Anti-Phishing Working Group, 2013, 2014). This escalating trend is mainly driven by the high profitability of the financial industry. In an analysis by the renowned brand protection company MarkMonitor Inc. (2015), the financial industry is found to be the most phished industry, accounting for 41% of the total phishing attacks in the first half of 2015.

The proliferation of phishing campaigns has discouraged consumers in using E-commerce websites, based on a recent survey on 1010 adults in the United States (American Institute of CPAs, 2015). The survey revealed that 86% of the participants are doubtful about the reliability of the businesses to safeguard their financial and other personal information. In terms of financial damage, it is estimated that worldwide organizations suffered losses amounting to $453 million in December 2014 (EMC Corporation, 2015). Besides inflicting heavy financial losses, phishing can also damage one's reputation (Purkait, De, & Suar, 2014). This claim can be observed in a unique phishing incident which happened in the United States (Timm & Perez, 2010). On January 21, 2009, Bryan Rutberg fell into a phishing scam. As a result, his Facebook account was taken over by phishers to post an emergency status. Shortly after that, Brian's friends on Facebook began receiving emails informing that he has been robbed while travelling in United Kingdom and desperately in need of financial aid to return home. One of Brian's friend responded by transferring $1200 to a Western Union bank

account in London, which belongs to the phishers. In this story, Brian himself did not suffer any loss of money. Instead, the phishers manipulate the compassion and trust of Brian's friends in order to steal their money.

In summary, phishing attacks have resulted in widespread privacy breaches and monetary loss, as well as shattering the reputation of individuals and businesses.

1.1 Problem Statement

This section highlights the major problems faced by existing anti-phishing methods. Although various anti-phishing methods have been introduced by security vendors and scholars, statistics have yet to show any substantial decline in the phishing crime rates. There is no fit-for-all method capable of providing complete protection against phishing, as phishers rapidly advance their strategies to deceive victims. An overview of the shortcomings in existing anti-phishing systems can be found in recent works by Zeydan et al. (2014) and Nirmal et al. (2015).

The first known weakness is the failure to catch newly launched (also known as zero-day) phishing webpages. This issue is most prevalent in conventional anti-phishing tools of the mainstream browsers such as Mozilla Firefox and Google Chrome (Google, 2015). The blacklist method works by checking each URL that a user intends to visit against a blacklist of previously reported phishing URLs. Although the blacklist method is most widely adopted, its inability to capture fresh phishing webpages presents a huge risk to the users.

Second, phishers continue to evolve their strategies by avoiding suspicious features in the phishing webpage (Wu, Du, & Wu, 2014). Hence, existing heuristic-based methods that capitalise on suspicious features are unstable and may become ineffective over the time.

Among the works that are based on heuristic approach includes Li et al. (2014), Nguyen et al. (2014) and Ramesh and Krishnamurthi (2014).

Third, language limitations are often found in techniques that utilise textual analysis. Existing works by Xiang and Hong (2009), Verma and Hossain (2014) and Ramesh et al. (2014) tend to rely on features and semantics that are exclusive to English language, thus making them not applicable for classifying non-English webpages (Zeydan et al., 2014).

Next, visual similarity-based methods such as Mao et al. (2013) and Chiew et al. (2015) can be vulnerable when the webpage layout or visual elements are updated or intentionally altered. In addition, it is costly to maintain an up-to-date database of legitimate images as a comparison reference. Thus, the incompleteness of the image database becomes a bottleneck to achieve good classification accuracy.

Lastly, identity-based methods tend to rely on the existence of legitimate URLs in the Hyper Text Markup Language (HTML) source codes in order to find out the target identity. Hence, the true positive rate in the methods proposed by Liu et al. (2010) and Ramesh et al. (2014) may plunge when the phishing webpage uses URLs that point to its own phishing domain name.

1.2 Research Objectives

To address the problems highlighted in previous section, the following research objectives are outlined:

(a) To investigate and identify the drawbacks of existing phishing detection methods.

(b) To propose and implement a novel and robust method that will automatically safeguard Internet users from becoming phishing victims.

(c) To improve the classification accuracy for non-English webpages.

1.3 Research Scope

This research focuses on phishing detection at webpage level, specifically at the instance when the user arrives at a phishing webpage. The proposed method can be deployed as a browser plug-in on the client-side to detect phishing webpages automatically and warn the users, while allowing legitimate webpages to be accessed normally.

1.4 Research Significance

This research carries the following significances:

(a) Effective detection of newly launched phishing webpages.

(b) Attain robustness in detecting phishing webpages hosted in any language.

(c) Offers long-term effectiveness by leveraging on permanent phishing characteristic.

(d) Enhance the confidence of consumers in using online transactions.

(e) Reduce financial losses faced by consumers and businesses.

1.5 Thesis Outline

This thesis is allocated into five distinct chapters, apart from references and appendices. The contents of the chapters are summarized as follows: