Efficient and Fast Server Based Phishing Detection System Using URL Lexical Analysis

by

AMMAR YAHYA DAEEF
(1540211762)

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy in Computer Engineering

School of Computer And Communication Engineering
UNIVERSITI MALAYSIA PERLIS

2017

UNIVERSITI MALAYSIA PERLIS

DECLARATION OF THESIS

Author's full name : AMMAR YAHYA DAEEF

Date of birth : 3-1-1984

Title : EFFICIENT AND FAST SERVER BASED PHISHING DETECTION SYSTEM USING URL LEXICAL ANALYSIS

Academic Session : 2017-2018

I hereby declare that the thesis becomes the property of Universiti Malaysia Perlis (UniMAP) and to be placed at the library of UniMAP. This thesis is classified as :

□ **CONFIDENTIAL** (Contains confidential information under the Official Secret Act 1972)*

□ **RESTRICTED** (Contains restricted information as specified by the organization where research was done)*

☑ **OPEN ACCESS** I agree that my thesis is to be made immediately available as hard copy or on-line open access (full text)

I, the author, give permission to the UniMAP to reproduce this thesis in whole or in part for the purpose of research or academic exchange only (except during a period of _____ years, if so requested above).

Certified by:

b/p

SIGNATURE **SIGNATURE OF SUPERVISOR**

A5606042 Prof. Ir. Dr. R Badlishah Ahmad

(NEW IC NO. / PASSPORT NO.) **NAME OF SUPERVISOR**

Date : 28/08/2017 Date : 28/08/2017

NOTES : * If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentially or restriction.

ACKNOWLEDGEMENTS

First and foremost, "All the praises and thanks be to Allah, The Beneficent, The Merciful"," praise be to Him who has taught by the pen, who taught man that which he knew not"

First of all, my sincere thanks to Allah, who endowed me to complete this PhD thesis. I would like to thank my supervisor, Prof. Ir. Dr. R Badlishah Ahmad, for all your guidance, support, brilliant ideas, patience, and the opportunities you have presented me. Your managerial skills and uncompromising quest for excellence always motivated me to present the best of what I can be.

I would like to express my heartfelt gratitude to my supervisor Dr. Yasmin Yacob for all numerous hours of discussions, support and encouragement. She has been helpful, understanding and generous throughout the study. She has truly been a mentor and I owe her my deepest thanks.

Finally, I would like to express my deep gratitude and thank to my beloved family for their love, patience and support, especially my beloved mother, my beloved brothers and sister for their continuous support and supplication. In addition, I would like to thank my dearest friends in Iraq for their support and prayers.

TABLE OF CONTENTS

CHAPTER 4 PRELIMINARY ANALYSIS FOR URL LEXICAL FEATURES

CHAPTER 5 MACHINE LEARNING CLASSIFIERS RESULTS

CHAPTER 6 CONCLUSION AND FUTURE WORK

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIWL	Automated Individual White List
AOL	America Online
APWG	Anti-Phishing Work Group
APWL	Anti-Phishing White List
ARFF	Attribute-Relation File Format
CCH	Contrast Context Histogram
DNS	Domain Name System
EMD	Earth Mover's Distance
FNR	False Negative Rate
FPR	False Positive Rate
HIPs	Human Interactive Proofs
HTTP	Hypertext Transfer Protocol
IP	Internet Protocol
LEO	Logo Extraction and Comparison
LM	Language Model
LMCL	Language Model based Classifier
LR	Logistic Regression
LUI	Login User Interface
MITB	Man in the Browser
ML	Machine Learning
MLE	Maximum Likelihood Estimation

NCL	National Consumers League
NGCL	N-gram based Classifier
NLP	Natural Language Processing
OA	Open Alexa
OCR	Optical Character Recognition
OD	Open DMOZ
PDOF	Phish Detect On Fly
PLSA	Probabilistic Latent Semantic Analysis
SC	Statistical Classifier
SIFT	Scale Invariant Feature Transform
SMS	Short Message Service
SPP	Single Password Protocol
SSL	Secure Sockets Layer
SVM	Support Vector Machine
TA	Tank Alexa
TCL	Token based Classifier
TD	Tank Alexa
TDF	Term Document Frequency
TIFF	Tagged Interchange File Format
UI	User Interface
URL	Uniform Resource Locator
US-CERT	The United States Computer Emergency Readiness Team

VoIP	Voice Over Internet Protocol

LIST OF SYMBOLS

Tokenphishratei	Token phish rate
URLphishrate	URL token phish rate
in f o(T)	The entropy function
C_j	URL class
T	set of choices
Split(T)	Information gain of split children
$K(x;x^{'})$	SVM kernel function
$h(x)$	The distance to the boundary of decision
$P(w)$	4-gram probability

Sistem Pengesan Memancing Data yang Cekap dan Pantas Menggunakan Pelayan Berdasarkan Analisis Leksikal URL

ABSTRAK

Pengesanan serangan phising ialah bidang penyelidikan yang signifikan untuk aplikasi keselamatan rangkaian. Laman web sahih selalunya terdedah kepada serangan phishing. Phishing menyebabkan cabaran berterusan dan terus menjadi ancaman menerusi pelbagai vector seperti enjin carian, laman web palsu, emel dan mesej segera. Penipuan berbentuk ini telah berevolusi untuk kekal satu langkah kehadapan oleh tindak balas terkini. Ia memanipulasi kelemahan pengguna yang menyebabkan penyelesaian masalah ini semestinya kompleks. Pengkelas phising menggunakan ekstrak fitur untuk mengesan laman phishing dan ia bergantung kepada sama ada kandungan laman web, Pengesan Sumber Seragam (URL) atau kedua-duanya. Pengekstrakan fitur URL mengandungi hos dan maklumat leksikal. Di dalam tesis ini pengekstrakan fitur hanya berdasarkan fitur leksikal untuk mengurangkan kos pemprosesan disebabkan oleh pengekstrakan fitur maklumat hos. Fitur-fitur ini digunakan oleh pengkelas untuk mengesan laman web phishing. Kebanyakan strategi pengesanan serangan phishing melayan mekanisme pengesanan pelanggan. Di dalam tesis ini, teknik baru pengesanan serangan phishing di cadang untuk mencapai sistem yang pantas, tegap dan tepat dengan menggunakan fitur leksikal sahaja. Bahagian pertama tesis mempersembahkan analisa dan pembangunan untuk fitur leksikal URL sedia ada termasuk tokenisasi dan mekanisme n-gram yang mengekstrak dan menganalisa token dan pengagihan n-gram yang sahih dan set data phishing diikuti dengan implementasi token berasaskan pengkelas (TCL) dan pengkelas beasaskan N-gram (NGCL). Oleh itu, TCL dan NGCL masing-masing memecahkan URL kepada token dan n-gram dan menggunakan pengagihan untuk proses klasifikasi. Juga, bahagian pertama tesis mencadangkan pengkelas berasaskan model bahasa (LMCL) yang membina model untuk kedua-dua kelas phishing dan sahih untuk mengklasifikasi URL berdasarkan kemungkinan tertinggi dan dibandingkan dengan pengkelas TCL dan NGCL. Bahagian kedua tesis mencadangkan penggunaan output LMCL sebagai pengkelas fitur tunggal yang digabungkan dengan fitur leksikal URL untuk membina fitur keseluruhan yang digunakan oleh pengkelas Mesin Pembelajaran (ML). Kemudian cadangan untuk meminda output LMCL untuk mengekstrak model fitur sub-bahasa dan menggabungkan dengan fitur leksikal URL untuk melatih pengkelas ML. Berikutan strategi ML bernama J48, pengkelas Mesin Sokongan Vektor (SVM) dan regresi logistic (LR) digunakan untuk mengesan URL phishing. Prestasi penilaian telah dicapai di semua peringkat untuk memenuhi pengesan serangan phishing yang pantas dan tepat. Sementara itu, kesemua pengkelas yang telah dicadangkan diuji menggunakan set data sebenar yang dikutip daripada pelbagai sumber untuk meneroka ketegaran teknik yang dicadangkan. Akhirnya, keputusan menunjukkan keboleharapan fitur leksikal berketepatan tinggi, tegar dan laju untuk pengesanan phishing URL. Diantara pengkelas yang dicadangkan, J48 dengan fitur cadangan menunjukkan keputusan keseluruhan terbaik dengan ketepatan 99% dan masa purata yang diperlukan untuk mengesan URL tunggal ialah 0.46 saat.

Efficient and Fast Server based Phishing Detection System Using URL Lexical Analysis

ABSTRACT

Phishing attack detection is a significant research area for network security applications. Legitimate websites is typically prone to phishing attacks. Phishing poses an ongoing challenge and continues to be a threat via numerous vectors such as search engines, fake websites, emails and instant messages. It has evolved its deceptions to remain one step ahead of the latest countermeasures. It exploits the weaknesses of the users which makes solving this problem especially complex. Phishing classifier uses the extracted features to detect the phishing websites and it depends on either the website's content, the Uniform Resource Locator (URL) or both of them. The URL feature extraction comprises host and lexical information. In this thesis, the feature extraction is based on the lexical features only in order to reduce the processing overhead due to the host information feature extraction. These features are utilized by a classifier to detect the phishing website. Most of the phishing attack detection strategies served the client side detection mechanisms. In this thesis, a new server side phishing attack detection technique is proposed to achieve fast, robust and accurate system by using lexical features alone. The first part of thesis presents analysis and development for the existing lexical features of URL including the tokenization and n-gram mechanisms which extract and analyze tokens and n-gram distribution of legitimate and phishing datasets followed by implementing Token based Classifier (TCL) and N-gram based Classifier (NGCL). Therefore, TCL and NGCL segment URLs into tokens and n-grams respectively and employ their distribution for classification process. Also, the first part of thesis proposing Language Model based Classifier (LMCL) which build a model for both of phishing and legitimate classes to classify URLs according to the highest probability and compared with TCL and NGCL classifiers. The second part of thesis proposing using the output of LMCL as a single classification feature in combination with URL lexical features in order to build the whole features that used by the Machine Learning (ML) classifiers. Then proposing to modify the output of LMCL to extract sub language model features and combined with URL lexical features to train ML classifiers. Regarding ML strategy J48, Support Vector Machine (SVM) and Logistic Regression (LR) classifiers are used for detecting the phishing URLs. The performance evaluation has been achieved regarding all these stages to meet a fast and accurate phishing attack detection. Meanwhile, all the proposed classifiers are tested using real life datasets collected from different sources in order to explore the robustness of the proposed techniques. Finally, the results showed a reliability of lexical features to provide high accuracy, robust and fast detection of phishing URLs. Among the proposed classifiers, J48 with the proposed features presents the best overall results with 99% accuracy and the time required to detect single URL is a 0.46 second on average.

CHAPTER 1

INTRODUCTION

The web has evolved widely in the life of people and since the beginning of Internet in 1990s, a lot of new security issues and threats appear continuously which constitute a challenge to users and security experts as well. Phishing is a cutting edge threat that has a deep impact on commercial and banking sectors by means of the Internet and delivers a huge misfortunes at the level of clients and organizations (Khonji, Iraqi, & Jones, 2013a). Phishing websites are highly similitude with the honest ones via trying to trap and bait users into these websites. Regarding this sort of attacks, phishers normally utilize technical and social designing traps together to begin their attacks. The social engineering attacks are focusing on users not on a system itself and intending to get the users information which are typically considered to be a touchy and confidential (Bozkir & Sezer, 2016).

Anti-Phishing Work Group (APWG)(Greg Aaron, 2016) reported that the number of phishing websites increased by 250% in the period from the last three months of 2015 to the first quarter of 2016 as shown in Fig. 1.1. The total number of discovered unique websites in the first quarter of 2016 is 289,371. Also, steadily rose per month was observed from October 2015 to March 2016 ranged from 48,114 to 123,555 respectively (Greg Aaron, 2016). These statistics demonstrate the significance to distinguish URLs and domain names to battle phishing. Additionally, the rise of online websites as highlighted in (Netcraft, 2016) reaches about one billion providing more services accessible on the Internet that can be targeted by phishers. Consequently, numerous new physical vectors, victims and targets get to be accessible letting space for a new sort of phishing attacks to be executed.

Different attack vectors are used to launch phishing attack such as search engines, fake websites, advisement, email, instant message, social media or phone call (G. Liu, Qiu, & Wenyin, 2010). This assortment of phishing attacks leads to a difficult protection against this phenomenon and existing phishing detection methods just adapting to a few of them. In spite of the broad field of phishing attack vectors, a typical purpose of numerous vectors is the utilization of the link misleading victims for phishing websites. Utilization of obfuscated Uniform Resource Locator (URL) and domain names are widely used in phishing attacks (Aaron & Rasmussen, 2015). URL obfuscation lures users by misleading them to forged websites via a URL or website of a genuine website familiar to the victim (Aaron & Rasmussen, 2015; Cova, Kruegel, & Vigna, 2008).

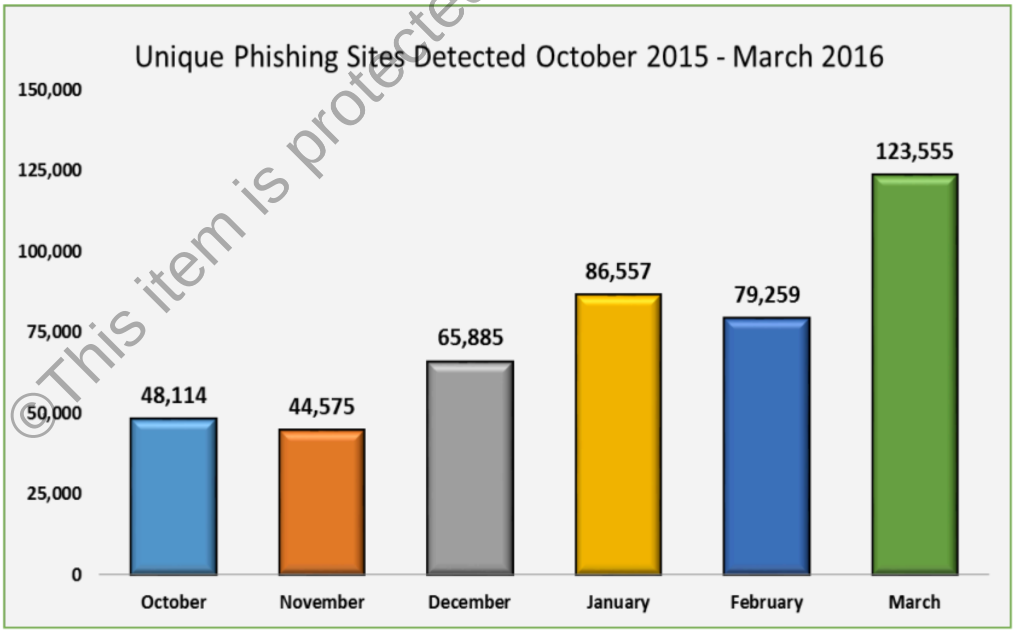

Figure 1.1: APWG phishing site trends 1st quarter 2016 (Greg Aaron, 2016).

Intelligent solutions based on phishing feature extraction (Abur-rous, Hossain, Dahal, & Thabtah, 2010; Almomani, Gupta, Atawneh, Meulenberg, & Almomani, 2013;

X. Chen, Bose, Leung, & Guo, 2011; Kazemian & Ahmed, 2015; Thomas, Grier, Ma, Paxson, & Song, 2011; Le, Markopoulou, & Faloutsos, 2011; J. Ma, Saul, Savage, & Voelker, 2011; Blum, Wardman, Solorio, & Warner, 2010a; J. Ma, Saul, Savage, & Voelker, 2009; Khonji, Iraqi, & Jones, 2011) depend on extracting important features of the website and after the extraction process these features utilized by an algorithm to decide or detect the phishing website. In general can be divided into content based and URL based solutions. Content based intercept and download the full contents of website for analyzing which can provide high detection accuracy with much more runtime overhead. In addition, it might accidentally provide more threats to users they look to keep safe from it. URL based techniques use a combination of host information and lexical features (Le et al., 2011; Thomas et al., 2011). Hosting information features need to be extracted from a remote server which in turn poses large latency to classify the URLs and prevent employing such methods for real time systems. While, URL lexical features are represented as bag-of-words result in huge vectors of features and cause processing overhead in addition to the low detection accuracy. Mostly, URL features are used to train a Machine Learning (ML) algorithms to generate a classifier to detect unseen URLs.

Generally, anti-phishing solutions can be positioned in different levels of attack flow where most researchers are focusing on client side solutions (Almomani et al., 2013; Khonji et al., 2013a; Heartfield & Loukas, 2016; Tewari, Jain, & Gupta, 2016; Aleroud & Zhou, 2017). The tools in client side include profile filter and browser toolbars. A few samples of a such tools can be specified by: CallingID[1], Spoof-Guard (Teraguchi &

[1] http://www.callingid.com/partners/safe-search/

Mitchell, 2004), IE phishing filter[2], NetCraft[3], CloudMark[4] and eBay toolbar[5]. However, client side tools add more processing overhead which can leads to lose the trust and satisfactory of users. Other factor that has always been challenging for the researcher and security expert in browser based techniques is the mode to display the warning messages. Passive warning used to notify about phishing, such as change in colour, pop-up with textual information displayed at the corner or periphery of browser without interrupting browse activity is either unnoticed or neglected by Internet user (Wu, Miller, & Little, 2006; Aleroud & Zhou, 2017; Zeydan, Selamat, & Salleh, 2014).

On the other hand, server side solutions are usually based upon approaches which use content filtering and form the best means of defending against zero-hour or zero-day phishing attempts. For this reason, most new developments to address zero-day attacks are based on server side applications (Khonji, Iraqi, & Jones, 2012). The server side filters and classifiers applications based on machine-learning techniques for phishing attack detection are divided into sub-sections such as bag-of-words model (Blanzieri & Bryl, 2008; A. Hamid & Abawajy, 2011; Wardman, Stallings, Warner, & Skjellum, 2011), multi classifiers algorithms (Miyamoto, Hazeyama, & Kadobayashi, 2008; Islam & Abawajy, 2013) , classifiers model based features (Islam & Abawajy, 2013; T.-C. Chen, Stepan, Dick, & Miller, 2014), clustering of phishing email (Bagirov, 2008; L. Ma, Yearwood, & Watters, 2009) and multi-layered system (Yearwood, Mammadov, & Banerjee, 2010; Olivo, Santin, & Oliveira, 2013; Abawajy & Kelarev, 2012). Generally, each has the same techniques, but has some differences in term of features extraction.

[2] https://support.microsoft.com/en-us/kb/930168

[3] http://toolbar.netcraft.com/

[4] http://www.cloudmarkdesktop.com/

[5] http://pages.ebay.co.uk/help/accounttoolbar-install.html

These previous studies of server side ML based techniques have built phishing classifiers to detect phishing websites using a combination of URL lexical features, hosting information, network traffic, and other strategies. Using lexical features only leads to low accuracy of detection which forces the designers of phishing classifiers to employ the other types of features such host information. Using such features required information to be looked up on a remote server. Though previous works had utilized URL lexical analysis as a component, what was lacking was the exploration of the full potential of a purely lexical approach to provide a high accurate and fast detection approach. Additionally, there is lack discussion of the delayed producing by these methods (A. Aggarwal, Rajadesingan, & Kumaraguru, 2012) and very little works stated the time required to detect a single URL which is considered unsuitable for real time application (Thomas et al., 2011; Le et al., 2011; Marchal, François, State, & Engel, 2014). As a consequence of restrictions in a current methods and the remembering that the most promising technique is URL analysis, especially the technique which depends only on lexical analysis and URL detection in a real time will be best familiar with minimum processing overhead. In addition, the main data entry points are usually a masqueraded URL (or link). Hence, this work proposing using URLs lexical features alone in order to explore the upper bound of performance can be achieved by URL lexical based phishing classifiers to provide high detection accuracy and minimum processing time to classify a single URL.

1.1 Problem Statement

Server side phishing detection systems are considered as the ideal solution to detect zero-day attacks online (Khonji et al., 2012; Almomani et al., 2013; Thomas et al., 2011). These systems should be lightweight enough to support the real time process and