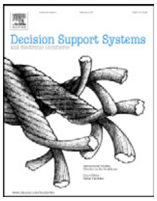
Decision Support Systems 88 (2016) 18–27

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss



# PhishWHO: Phishing webpage detection via identity keywords extraction and target domain name finder

Choon Lin Tan[a], Kang Leng Chiew[a,*], KokSheik Wong[b], San Nah Sze[a]

[a] Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
[b] Faculty of Computer Science and Information Technology,University of Malaya, 50603 Kuala Lumpur, Malaysia

ARTICLE INFO

*Article history:*
Received 20 October 2015
Received in revised form 28 March 2016
Accepted 22 May 2016
Available online 1 June 2016

*Keywords:*
Phishing detection
Identity keywords
N-gram
Weighted URL tokens
Search engine

ABSTRACT

This paper proposes a phishing detection technique based on the difference between the target and actual identities of a webpage. The proposed phishing detection approach, called PhishWHO, 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 is proposed to determine the legitimacy of the query webpage. The overall experimental results suggest that the proposed system outperforms the conventional phishing detection methods considered.

## 1. Introduction

In this modern age of information technology, consumers are dealing with more products and services through the online channel. Therefore, having multiple online accounts (e.g., email account, banking account, social networking account) have become a norm for most people. This technological trend is exposing internet users to a rising threat of online identity theft known as phishing [17].

Phishing websites are counterfeit websites designed to deceive victims and steal their account login credentials, credit card numbers or other personal secrets. Phishers usually entice victims to the phishing website by sending emails containing the fraudulent URL and some threatening messages such as possible account termination, and fake alert on illegal transaction [9]. At the phishing website, the phishers will capture sensitive information submitted by the victims.

The severity of phishing threats in recent years continues to escalate, based on statistics gathered from security organizations. For instance, a total of 42,212 unique phishing websites was reported in June 2014 by the Anti-Phishing Working Group [2], whereas the financial loss inflicted upon worldwide organization in December 2014 was estimated to be $453 million [10]. These alarming trends have resulted in the loss of consumers' trust in using E-commerce websites because they are feared to become fraud victims [6]. In summary, phishing attacks have resulted in widespread leakage of sensitive information, monetary loss and crippled businesses' reputation.

The key factor that makes phishing possible is the human behaviour when interacting with electronic communication channels. Dhamija et al. [8] identified several user tendencies that are exploited by phishing attacks. For instance, a typical user is often unaware of the significance of common security indicators such as the Secure Sockets Layer (SSL) icon and digital certificate on the browser address bar. As a result, these useful indicators are often ignored. In addition, some users are confused on how a legitimate URL is supposed to resemble, thus they rely on the webpage contents to determine its genuineness [18]. A recent assessment by Alsharnouby et al. [1] reveals that participants with phishing awareness can only achieve 53% of average success rate in identifying phishing websites. These studies have proven that both normal and technical users can be easily deceived by phishing webpages. Hence, it is crucial to have an efficient phishing detection system, where users can be effectively safeguarded from phishing attacks.

To compensate for the human limitations in detecting phishing websites, automated solutions have been introduced in conventional web browsers and security applications. Most solutions rely on blacklists (e.g., Google Safe Browsing list, PhishTank list) that

[*] Corresponding author. Tel.: +6082 58 3762.
*E-mail addresses:* colin89lin@gmail.com (C. Tan), klchiew@unimas.my (K. Chiew), koksheik@um.edu.my (K. Wong), snsze@unimas.my (S. Sze).

http://dx.doi.org/10.1016/j.dss.2016.05.005
0167-9236/© 2016 Elsevier B.V. All rights reserved.