Building a Frame-Based Anti-Phishing Model based on Phishing Ontology

Shian-Shyong Tseng
Dept. of Applied Informatics and Multimedia, Asia University, Wufeng, Taiwan
sstseng@twnic.net.tw

Ching-Heng Ku
Taiwan Network Information Center, Taipei, Taiwan
chku@twnic.net.tw

Tsug-Ju Lee
Dept. of Applied Informatics and Multimedia, Asia University, Wufeng, Taiwan
freeman1217@gmail.com

Guang-Gang Geng
China Internet Network Information Center, Computer Network Information Center, Chinese Academy of Sciences, Beijing, China
gengguanggang@cnnic.cn

Yuh-Jye Wang
Dept. of Applied Informatics and Multimedia, Asia University, Wufeng, Taiwan
yjwang@vghtc.gov.tw

Abstract

In recent years, with the rapid growth of the Internet applications and services, Phishing attacks seriously threaten the web security. Generally, anti-phishing methods either use blacklists or recognize the phishing pattern with statistical learning. Due to the versatile and dynamic nature of phishing patterns, the development and maintenance of the anti-phishing prevention system is difficult and their cost is expensive. Hence, how to acquire and update the phishing knowledge in the anti-phishing detection system become an important issue. In this study, we use the frame-based approach to build up an anti-phishing attack model where the frame is used to model the phishing attack scenario and knowledge with the proposed phishing ontology. Finally, the prototype of the frame-based anti-phishing model has been successfully built to show the feasibility in the detection of the phishing attack behavior.

Keywords: Frame-based anti-phishing model, Phishing attack detection, Phishing ontology

1. INTRODUCTION

Internet Phishing attack has become the fastest growing scam on the Internet. Phishers usually cheat Internet users of their credit card number, password or personal sensitive information by utilizing various Internet mediums such as email, spoofed websites or advertisements [1].

According to the 2012 annual report of Anti-Phishing Alliance of China (APAC) [2], APAC handled 24,535 phishing websites in 2012. The distribution of phishing websites remain mainly in payment/transaction, finance/securities and media/communication websites or pages that involve online login and payment.

Along with an increase in the number of potential targets, three major factors [3] including unawareness of threat, unawareness of policy, and criminal's technical sophistication have been utilized by criminals to take advantage. They not only use fake email messages and web sites to lure users into divulging their personal information, but also increasingly use malicious codes that specifically target user account information. Furthermore, phishers today have a large tackle box of tools available to them[3]. Some phishing attacks usually use compound tricks and the unawareness of the attack that become more and more difficult for users [6].

In general, anti-phishing strategies can be categorized into silently eliminating the threat, warning users about the threat and training users not to fall for attacks. In response to these threats, anti-phishing researchers have developed various solutions [8-9] for the anti-phishing. In order to find out a suspicious webpage, researchers must first identify the legitimate webpage under attack — that is, the phishing target[4]. Unfortunately, this requirement isn't always easy to satisfy for general scenarios. Some different approaches [5, 10-13] on the feature extraction had been proposed for the phishing webpage detection. Moreover, the DNS based[14], the collaborative database[15], the decisive heuristics[16], and the genetic algorithm based[17] approaches had also been proposed for the anti-phishing. Otherwise, Hossam[18] proposed a

hybrid anti-phishing tool. Thiyagarajan[19] and Antonio[20] used the anti-phishing approaches on the e-banking services.

However, phishing attacks have become increasingly scale and industrialized trends. Existing phishing detection tool still suffer false alarms and false negatives. Thus, responding to the changing phishing attack knowledge is very important.

In this study, we focus on the study of phishing detection model that can be used to detect phishing pages. We intend to proceed from the point of view of the phishing attacks to study the patterns of the behavior of the attacker to make the phishing detection and large-scale fishing warning can be achieved when the frame-based anti-phishing model is established related to the targeted web site. The contribution of this study firstly is to make the control flow and phishing pattern separately in the intelligent anti-phishing system to achieve the toleration of the possible phishing pattern in the future. The system using the proposed phishing ontology can generalize the phishing behaviour for the extraction of the phishing knowledge. Secondly, the proposed frame-based anti-phishing model can concentrate on the phishing scenario and phishing knowledge for the targeted webpage with the sensitive information to increase the efficiency of the anti-phishing detection. Therefore, the inheritance and instantiation properties of the frame model allow us to be able to easily extend or update the phishing prevention scenarios to increase the detection of the phishing attack.

In the following section, we will introduce the phishing ontology and the frame-based anti-phishing model in Sections 2 and 3, respectively. The experiment will be shown in the Section 4. The conclusion is described in the last section.

2. PHISHING ONTOLOGY

An ontology that used in research on Artificial Intelligence and Knowledge Representation is defined as an explicit specification of a conceptualization by Tom Gruber [22]. For knowledge-based systems, what "exists" is exactly that which can be represented. When the knowledge of a domain is represented in a declarative formalism, the set of objects that can be represented is called the universe of discourse. This set of objects, and the describable relationships among them, are reflected in the representational vocabulary with which a knowledge-based program represents knowledge. Thus, we can describe the ontology of a program by defining a set of representational terms. In such an ontology, definitions associate the names of entities in the universe of discourse (e.g., classes, relations, functions, or other objects) with human-readable text describing what the names are meant to denote, and formal axioms that constrain the interpretation and well-formed use of these terms.

According to Guarino's categorization[23], the conception of the ontology can be categorized as three parts, such as Terminological ontologies, Information ontologies, and knowledge modeling ontologies. Besides, the ontology defined in the system can be composed of the three concepts of entities, attributes, and associated relations.

Because the Phishing knowledge need to be increased according to the practical environment, hence, how to systematically categorize the anti-phishing knowledge is an important issue. In this study, the Bloom's taxonomy of cognitive domain was used to classify anti-phishing knowledge into different levels of learning skills. Figure 1 shows the proposed phishing ontology. The Bloom's taxonomy of cognitive domain is used to define different level of anti-phishing learning concepts. Concepts in higher levels, such as analysis and application levels, are more general and can be used as principles of network literacy. Concepts in lower levels, such as comprehension and knowledge levels, are more specific anti-phishing patterns for memorizing.

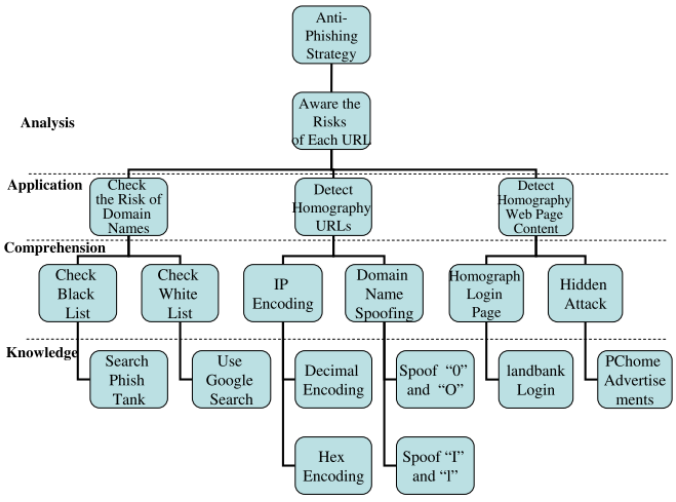

Figure 1. Phishing Ontology leveled by Bloom's Taxonomy

Based on the phishing ontology, various anti-phishing cases can be categorized and labeled by specific anti-phishing strategy concept. Thus, the

corresponding interactions and phishing knowledge can be proposed.

In this study, the proposed ontology of the phishing knowledge includes a set of attributes related to the web frame and the relationships among those frames described in the following section.

3. FRAME-BASED ANTI-PHISHING MODELING

We observe that the phishing page can be composed of two components, such as the phishing knowledge and the page scenario. Phishing knowledge stands for the attack technique which is used to spoof users. Page scenario is where the attacks take place. Different composition phishing knowledge and page scenario will lead to different phishing cases. For example, a spoofed yahoo page with the phishing URL "http://www.yahO0.com.tw" is the composition of the portal page of yahoo, and phishing attack trick of replacing alphabet "O" with zero "0".

In this paper, the frame on Phishing attack knowledge and Phishing scenario can be simply described by stereotyped attributes. We represent Phishing attack scenario and Phishing attack knowledge as scenario frame (SF) and knowledge frame (KF) by analysing the common Phishing attack cases related to the proposed phishing ontology in the previous section. Besides, the inheritance and instantiation properties of our proposed frame-based knowledge representation can further extend the original pages to either new attack or new scenario. Thus, a large amount of new scenario pages can be generated. The detail is described in the following section.

Our frame-based approach contains two main parts, Phishing attack scenario frame (SF) is used to represent Phishing attack scenarios and Phishing attack knowledge frame (KF) is used to manage Phishing attack techniques. The detail is described in the following.

3.1. Phishing attack scenario frame (SF)

A Phishing attack scenario frame (SF) is created by the common Phishing case with multiple scenes. In Fig.2, the scenario consists of three scenario pages, P01, P02, and P03. Attacker defrauds victim with malicious e-mail and interesting topic in P01. Then, users login the web site with obfuscated URL by username and password in P02. Finally, hacker acquires information of username and password and then redirect user to real web site in P03. Once, users carelessly check visiting page legitimate or the target of desiring hyperlink, their sensitive private information will suffer from great dangerous of leaking. The attack is called "URL obfuscation". The objects representing the items in scenario page are also defined in scenario frame. In this study, we transform the mail, web action form, and URL address as objects, M1, F1, and U1, respectively, to represent frames.

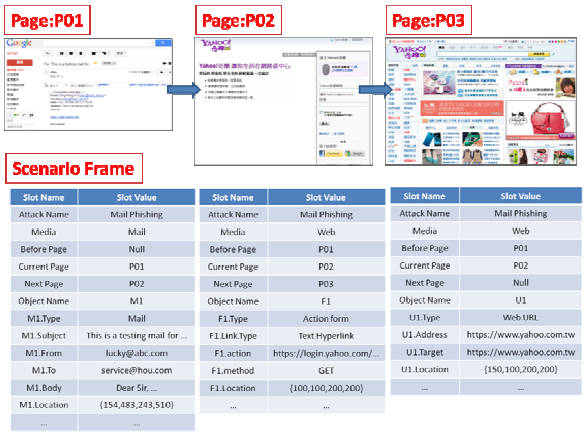

Figure 2. Frame-based Anti-Phishing Scenario

3.2. Phishing attack knowledge frame (KF)

In our frame-based approach, we acquire Anti-Phishing attack knowledge from domain experts to construct the knowledge frame (KF) from the ontology. Below is an example of Phishing attack knowledge frame (KF). The attack knowledge can be represented as frame with six stereotype attributes of Phishing attack knowledge frame. "Security Issue" is the title of attack. "Security Issue No" is the knowledge number. "Media" means the container of Phishing attack, such as email, web, and mobile. "Suit To" represents scenario pages that are associated to attack knowledge. "Issue Description" describes the attack with a short description. "Action Script" is script language used for object operation. We can use action script to manipulate the object value in scenario pages. Fig.3 shows the slot attributes of attack knowledge frame describing the case of "URL obfuscation". This Phishing attack replaces the similar letter in page content appearance and similar URL appearance to cheat users. For example, in knowledge K026 and K027, the letter "O" can be replaced by the Arabic numberal "0"

and the letter "l" can be replaced by the Arabic numberal "1".

Slot Name	Slot Value
Security Issue	Homography with similar word (o, O, 0)
Security Issue No	K026
Media	Web, Mobile, Mail
Suit To	S01, S02, S03, ...
Issue Description	Homography by replace with similar word.
Action Script	If Object.String contain "o" then replace with "0". If Object.String contain "O" then replace with "0". If Object.String contain "0" then replace with "o". If Object.String contain "0" then replace with "O".
...	...

(a)

Slot Name	Slot Value
Security Issue	Homography with similar word (I, l, 1)
Security Issue No	K027
Media	Web, Mobile, Mail
Suit To	S01, S02, S03, ...
Issue Description	Homography by replace with similar word.
Action Script	If Object.String contain "l" then replace with "1". If Object.String contain "l" then replace with "1". If Object.String contain "1" then replace with "l". If Object.String contain "1" then replace with "l".
...	...

(b)

Figure 3. Examples of Phishing attack knowledge frame. (a) the letter "O" replaced by the Arabic numberal "0". (b) the letter "l" replaced by the Arabic numberal "1".

3.3. Phishing scenario extensions with knowledge frame

Based upon the features of frame-based approach, the new anti-phishing attack knowledge can be produced easily with frame format. The frame-based format makes the knowledge extension easily. There are two extension methods. First, we can extend the malicious anti-phishing attack scenario content by combining the scenario frame (SF) and knowledge frame (KF). According to action script in KF, we can modify the URL by replace similar letter in URL address. A new scenario content can be generated with a malicious F1.action URL "http://www.yah00.com.tw". This is a "URL obfuscation" attack. For a given phishing attack knowledge, we can apply it on all the scenario pages satisfying the precondition of phishing attack knowledge. This instantiation property provides us another way to keep the extensibility of contents from the inheritance property.

Second, we can enhance the current knowledge base by integrating the new attack knowledge. For example, two Phishing attack knowledge frames (KFs) can be integrated into a new Phishing attack knowledge frame (KF). It can be considered as a new Phishing attack knowledge using two Phishing attack techniques.

4. EXPERIMENT

We use the frame-based approach to build up the anti-phishing model where the frame is used to model the phishing attack scenario and knowledge. Therefore, the inheritance and instantiation properties of the frame model allow us to be able to easily extend or update the phishing prevention scenarios to increase the efficiency of the phishing detection.

The proposed model has been applied in the Phishing attack case that contains three pages, P01, P02, and P03, as shown in Fig.4.

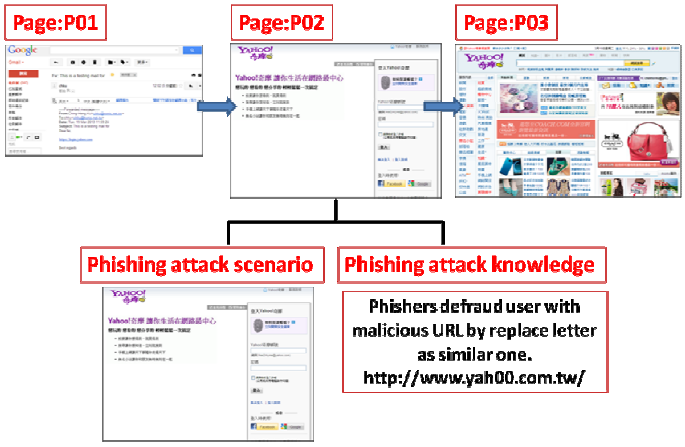

Figure 4. The Decomposition of a phishing attack case

We find out that the F1.action in scenario frame contains URL http://www.yahoo.com.tw, as shown in Fig. 5.

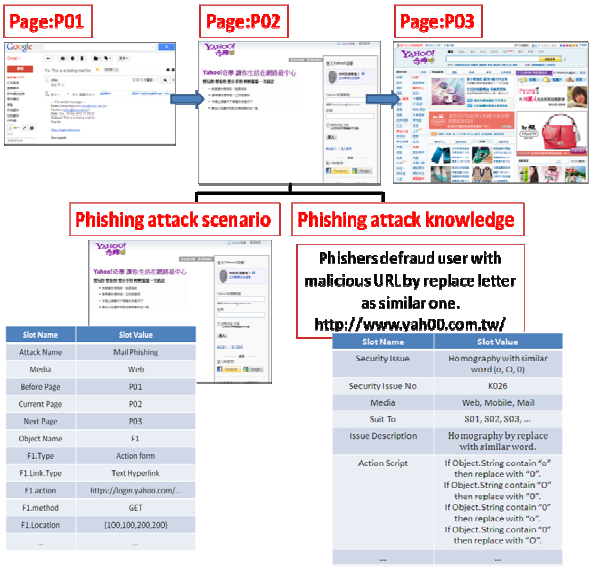

Figure 5. Scenario content generation using SF and KF

In order to represent the characteristics of the objects embedded in scenario page, each type of object has its own specific attributes. As shown in Fig.6, the object F1 represents the web action form with the frame attributes including type, link type, action, method, and location. In Fig.7, the example object S1 represents the SMS of mobile device with attributes of object S1 including type, from number, to number, date, message, and location.

Figure 6. An example of the "ACTION FORM" object in Web application

Figure 7. An example of the "SMS" message object in mobile phone

We integrate different attack knowledge to generate new attack knowledge. It can be considered as the new Phishing attack knowledge implemented by these two Phishing attack techniques.

In Fig.6, for example, "Similar character attack on hyperlink: O and 0" is to create spoofed URL by replace "O" in URL by "0"; and "Graph substitution attack" is to maintain the same link appearance by embedding graph file and the corresponding underlying link target leads to different URL of Phishing scenario. These two Phishing attack techniques can be combined into "Graph substitution attack with similar character attack on link target: O and 0". It is the knowledge fusion rule.

5. CONCLUSIONS

In this study, we use the frame-based approach to build up the anti-phishing attack model where the frame is used to model the phishing attack scenario and knowledge with the proposed phishing ontology. We make the control flow and phishing pattern separately in the intelligent anti-phishing system to achieve the toleration of the possible phishing pattern in the future. The system using the proposed phishing ontology can generalize the phishing behaviour for the extraction of the phishing knowledge. Besides, the proposed frame-based anti-phishing model can concentrate on the phishing scenario and phishing knowledge for the targeted webpage with the sensitive information to increase the efficiency of the anti-phishing detection. Finally, the frame-based anti-phishing model has been successfully built to show the feasibility of the detection of the phishing attack behavior.

ACKNOWLEDGMENT

This paper is partially sponsored by the Taiwan Network Information Center (TWNIC) and the China Internet Network Information Center (CNNIC) under the number of DNSLAB-2012-S-U.

REFERENCES

[1] Phishing Scams: Understanding the latest trends, June 2004.

[2] 2012 annual report of Anti-Phishing Alliance of China(APAC), 2012, http://en.apac.cn/news/201301/P020130122639769507177.pdf

[3] Jason Milletary, "Technical trends in Phishing attacks", http://www.us-cert.gov/sites/default/files/publications/phishing_trends0511.pdf, US-CERT.

[4] W. Liu et al., "An Antiphishing Strategy Based on Visual Similarity Assessment," *IEEE Internet Computing*, vol. 10, no. 2, 2006, pp. 58–65.

[5] Liu Wenyin, Gang Liu, Bite Qiu, and Xiaojun Quan, "Antiphishing through Phishing Target Discovery", *Proceedings of the IEEE Internet Society*, pp. 52-60, 2012.

[6] Tsung-Ju Lee, et al. "Game-based Anti-Phishing Training", *Proceedings of the TWELF*, 2010.

[7] P. Kumaraguru, et al. "Teaching Johnny not to fall for phish", *Proceedings of the ACM Transaction on Internet Technology*, 2007.

[8] Ming Qi and Chang-Yi Zou, "A study of anti-phishing strategies based on TRIZ", *Proceedings of the International Conference on Networks Security, Wireless Communications and Trusted Computing*, pp. 536-538, 2009.

[9] Carly Wilson and David Argles, "The fight against phishing: technology, the end user and legislation", *Proceedings of IEEE*, pp. 501-503, 2011.

[10] Liu Wenyin, etc., "Phishing webpage detection", *Proceedings of the Eight International Conference on Document Analysis and Recognition(ICDAR'05)*, 2005.

[11] Chun-Ying Huang, etc., "Mitigate web phishing using site signatures", *Proceedings of the IEEE TENCON 2010*, pp. 803-808, 2010.

[12] Hossain Shahriar and Mohammad Zulkernine, "Information source-based classification of automatic phishing website detectors", *Proceedings of the IEEE/IPSJ International Symposium on Applications and the Internet*, pp. 190-195, 2011.

[13] Weiwei Zhuang, etc., "An intelligent anti-phishing strategy model for phishing website detection", *Proceedings of the 32nd International Conference on Distributed Computing Systems Workshops*, pp. 51-56, 2012.

[14] Sun Bin, etc., "A DNS based anti-phishing approach", *Proceedings of the Second International Conference on Networks Security, Wireless Communications and Trusted Computing*, pp. 262-265, 2010.

[15] Arash Nourian, etc., "CASTLE: A social framework for collaborative anti-phishing databases", 2009. http://dx.doi.org/10.4108/ICST.COLLABORATECOM2009.8310

[16] Sophie Gastellier-Prevost, etc., "Decisive heuristics to differentiate legitimate from phishing sites", *Proceedings of IEEE*, 2011.

[17] V. Shreeram, etc., "Anti-phishing detection of phishing attacks using genetic algorithm", Proceedings of ICCCCT'10, pp. 447-450, 2010.

[18] Hossam M.A.Fahmy and Salma A.Ghoneim, "PhishBlock: A hybrid anti-phishing tool", 2009

[19] Thiyagarajan P, etc., "Anti-phishing technique using automated challenge response method", *Proceedings of the International Conference on Communication and Computational Intelligence*, pp. 585-590, India, December 2010.

[20] Antonio San Martino and Xavier Perramon, "Defending E-Banking services: antiphishing approach", *Proceedings of the Second International Conference on Emerging Security Information, Systems and Technologies*, pp. 93-98, 2008.

[21] Koen Kerremans, etc.,"Towards Ontology-based E-mail Fraud Detection", *Proceedings of the Portuguese Conference on Artificial intelligence*, pp.106-111, 2005.

[22] Tom Gruber, "A translation approach to portable ontology specifications", Knowledge Acquisition, pp. 199-220, 1993.

[23] Guarino, N., "Formal Ontology and Information Systems", *Proceedings Of the 1st International Conference, Trento*, Italy, pp.3-15, 1998.