Expert Systems with Applications 37 (2010) 7913–7921

Contents lists available at ScienceDirect

Expert Systems with Applications

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

# Intelligent phishing detection system for e-banking using fuzzy data mining

Maher Aburrous [a,*], M.A. Hossain [a], Keshav Dahal [a], Fadi Thabtah [b]

[a] Department of Computing University of Bradford, Bradford, UK
[b] MIS Department Philadelphia University Amman, Jordan

ARTICLE INFO

*Keywords:*
Phishing
Fuzzy logic
Data mining
Classification
Association
Apriori
E-banking risk assessment

ABSTRACT

Detecting and identifying any phishing websites in real-time, particularly for e-banking, is really a complex and dynamic problem involving many factors and criteria. Because of the subjective considerations and the ambiguities involved in the detection, fuzzy data mining techniques can be an effective tool in assessing and identifying phishing websites for e-banking since it offers a more natural way of dealing with quality factors rather than exact values. In this paper, we present novel approach to overcome the 'fuzziness' in the e-banking phishing website assessment and propose an intelligent resilient and effective model for detecting e-banking phishing websites. The proposed model is based on fuzzy logic combined with data mining algorithms to characterize the e-banking phishing website factors and to investigate its techniques by classifying the phishing types and defining six e-banking phishing website attack criteria's with a layer structure. Our experimental results showed the significance and importance of the e-banking phishing website criteria (URL & Domain Identity) represented by layer one and the various influence of the phishing characteristic on the final e-banking phishing website rate.



## 1. Introduction

E-banking phishing websites are forged websites that are created by malicious people to mimic real e-banking websites. Most of these kinds of web pages have high visual similarities to scam their victims. Some of these web pages look exactly like the real ones. Unwary Internet users may be easily deceived by this kind of scam. Victims of e-banking phishing websites may expose their bank account, password, credit card number, or other important information to the phishing web page owners. The impact is the breach of information security through the compromise of confidential data, and the victims may finally suffer losses of money or other kinds. Phishing is a relatively new Internet crime in comparison with other forms, e.g., virus and hacking. More and more phishing web pages have been found in recent years in an accelerative way (Fu, Wenyin, & Deng, 2006). The word phishing from the phrase "website phishing" is a variation on the word "fishing". The idea is that bait is thrown out with the hopes that a user will grab it and bite into it just like the fish. In most cases, bait is either an e-mail or an instant messaging site, which will take the user to hostile phishing websites (James, 2006).

E-banking phishing website is a very complex issue to understand and to analyze, since it is joining technical and social problem with each other for which there is no known single silver bullet to entirely solve it. The motivation behind this study is to create a resilient and effective method that uses fuzzy data mining algorithms and tools to detect e-banking phishing websites in an automated manner.

DM approaches such as neural networks, rule induction, and decision trees can be a useful addition to the fuzzy logic model. It can deliver answers to business questions that traditionally were too time-consuming to resolve such as, "Which are most important e-banking phishing website characteristic indicators and why?" by analyzing massive databases and historical data for training purposes.

The paper is organized as follows: Section 2 presents the literature review and related work, Section 3 shows the theory and methodology of the proposed fuzzy based data mining approach for the phishing website risk assessment model. Section 4 introduces the system design and implementation with the overall fuzzy data mining inference rules. Section 5 reveals the experiments and results of the fuzzy data mining e-banking phishing website risk assessment model and then conclusions and future work are given in Section 6.

## 2. Literature review and related work

### 2.1. Literature review

Phishing website is a recent problem, nevertheless due to its huge impact on the financial and on-line retailing sectors and since

* Corresponding author.
*E-mail addresses:* mrmaburr@bradford.ac.uk (M. Aburrous), m.a.hossain1@bradford.ac.uk (M.A. Hossain), k.p.dahal@bradford.ac.uk (K. Dahal), ffayez@philadelphia.edu.jo (F. Thabtah).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.04.044

preventing such attacks is an important step towards defending against e-banking phishing website attacks, there are several promising defending approaches to this problem reported earlier. In this section, we briefly survey existing anti-phishing solutions and list of the related works. One approach is to stop phishing at the e-mail level (Adida, Hohenberge, & Rivest, 2005), since most current phishing attacks use broadcast e-mail (spam) to lure victims to a phishing website (Wu, Miller, & Garfinkel, 2006a). Another approach is to use security toolbars. The phishing filter in IE7 (Sharif, 2006) is a toolbar approach with more features such as blocking the user's activity with a detected phishing site. A third approach is to visually differentiate the phishing sites from the spoofed legitimate sites. Dynamic Security Skins (Dhamija & Tygar, 2005) proposes to use a randomly generated visual hash to customize the browser window or web form elements to indicate the successfully authenticated sites. A fourth approach is two-factor authentication, which ensures that the user not only knows a secret but also presents a security token (FDIC, 2004). However, this approach is a server-side solution. Phishing can still happen at sites that do not support two-factor authentication. Sensitive information that is not related to a specific site, e.g., credit card information and SSN (Social Security Number), cannot be protected by this approach either (Wu, Miller, & Little, 2006b).

Many industrial anti-phishing products use toolbars in web browsers, but some researchers have shown that security tool bars do not effectively prevent phishing attacks. Bridges and Vaughn (2001), Dhamija and Tygar (2005) proposed a scheme that utilises a cryptographic identity-verification method that lets remote web servers prove their identities. However, this proposal requires changes to the entire web infrastructure (both servers and clients), so it can succeed only if the entire industry supports it. In Liu, Deng, Huang, & Fu (2006), the authors proposed a tool to model and describe phishing by visualizing and quantifying a given site's threat, but this method still would not provide an anti-phishing solution. Another approach is to employ certification, e.g., Microsoft spam privacy (Microsoft, 2004; Microsoft Corp., 2005; Olsen, 2004; Perez, 2003; WholeSecurity Web Caller, 2005). A recent and particularly promising solution was proposed in Herzberg and Gbara (2004), which combines the technique of standard certificates with a visual indication of correct certification; a site-dependent logo indicating that the certificate was valid would be displayed in a trusted credentials area of the browser. A variant of web credential is to use a database or list published by a trusted party, where known phishing websites are blacklisted. For example, Netcraft anti-phishing toolbar (Netcraft, 2004) prevents phishing attacks by utilising a centralized blacklist of current phishing URLs. Other examples include websense, McAfee's anti-phishing filter, Netcraft anti-phishing system, Cloudmark SafetyBar, and Microsoft Phishing Filter (Pan & Ding, 2006). The weaknesses of this approach are its poor scalability and its timeliness. Note that phishing sites are cheap and easy to build and their average lifetime is only a few days. APWG provides a solution directory at Anti-Phishing Working Group (2007) which contains most of the major anti-phishing companies in the world. However, an automatic anti-phishing method is seldom reported. The typical technologies of anti-phishing from the user interface aspect are done by Dhamija and Tygar (2005) and Wu et al., 2006b. They proposed methods that need web page creators to follow certain rules to create web pages, either by adding dynamic skin to web pages or adding sensitive information location attributes to HTML code. However, it is difficult to convince all web page creators to follow the rules (Fu et al., 2006). In Fu et al. (2006), Liu, Huang, Liu, Zhang, & Deng, 2005; Liu et al., 2006), the DOM-based (Wood, 2005) visual similarity of web pages is oriented, and the concept of visual approach to phishing detection was first introduced. Through this approach, a phishing web page can be detected and reported in an automatic way rather than involving too many human efforts. Their method first decomposes the web pages (in HTML) into salient (visually distinguishable) block regions. The visual similarity between two web pages is then evaluated in three metrics: block level similarity, layout similarity, and overall style similarity, which are based on the matching of the salient block regions (Fu et al., 2006).

### 2.2. Main characteristics of e-banking phishing websites

Evolving with the anti-phishing techniques, various phishing techniques and more complicated and hard-to-detect methods are used by phishers. The most straightforward way for a phisher to defraud people is to make the phishing web pages similar to their targets. Actually, there are many characteristics and factors that can distinguish the original legitimate website from the forged e-banking phishing website like Spelling errors, Long URL address and Abnormal DNS record. The full list is shown in Table 1 which will be used later on our analysis and methodology study.

### 2.3. Why using fuzzy logic and data mining?

FL has been used for decades in the engineering sciences to embed expert input into computer models for a broad range of

**Table 1**
Components and layers of e-banking phishing website criteria.

| Criteria | *N* | Component | Layer No. |
|---|---|---|---|
| URL & Domain Identity (Weight = 0.3) | 1 | Using the IP address | Layer 1 |
| | 2 | Abnormal request URL | |
| | 3 | Abnormal URL of Anchor | |
| | 4 | Abnormal DNS record | Sub weight = 0.3 |
| | 5 | Abnormal URL | |
| Security & Encryption (Weight = 0.2) | 1 | Using SSL certificate | Layer 2 |
| | 2 | Certification authority | |
| | 3 | Abnormal Cookie | |
| | 4 | Distinguished Names Certificate (DN) | |
| Source Code & Java script (Weight = 0.2) | 1 | Redirect pages | |
| | 2 | Straddling attack | |
| | 3 | Pharming attack | Sub weight = 0.4 |
| | 4 | Using onMouseOver to hide the Link | |
| | 5 | Server Form Handler (SFH) | |
| Page Style & Contents (Weight = 0.1) | 1 | Spelling errors | Layer 3 |
| | 2 | Copying website | |
| | 3 | Using forms with "Submit" button | |
| | 4 | Using Pop-Ups windows | |
| | 5 | Disabling Right-Click | |
| Web Address Bar (Weight = 0.1) | 1 | Long URL address | Sub weight = 0.3 |
| | 2 | Replacing similar characters for URL | |
| | 3 | Adding a prefix or suffix | |
| | 4 | Using the @ Symbol to Confuse | |
| | 5 | Using Hexadecimal Character Codes | |
| Social Human Factor (Weight = 0.1) | 1 | Much emphasis on security and response | |
| | 2 | Public generic salutation | |
| | 3 | Buying Time to Access Accounts | |
| | Total weight | | 1 |