AN ANTI-PHISHING MODEL FOR ECOMMERCE UNDER A NETWORK ENVIRONMENT

Yifei Cheng and Gen Li
School of Economics and Management, Harbin Engineering University, Nantong St.145-11, Harbin City, China

Keywords: Phishing, Ecommerce, Fuzzy algorithm.

Abstract: It is a complicated problem to detect the ecommerce phishing websites in real time. The detection is affected by a lot of factors which are indistinct and correlative. Therefore, fuzzy classification tools can translate the phishing websites details into crisp numbers which could be understood by human. In this paper, we proposed a model which includes six standards and twenty seven elements with fuzzy data mining method to detect and assess the ecommerce phishing risk rate. Although lack of the specific data to do experiment to check the validity, it is a good start to focus on the ecommerce security.

1 INTRODUCTION

The word phishing is abbreviation of the phrase “website phishing”. The idea is that bait is launched and to wish a victim will bite it just like a fish. When a user is browsing the B2C or C2C websites (e.g. Taobao, Youa, Alibaba), the criminals may send an e-mail or a link URL from instant messaging services such as QQ or other web messengers which will take the user to hostile phishing websites (James, 2006). At present, phishing websites are overflow with the rapid development of ecommerce which are fabricated to feign the real ecommerce websites by criminals. With the advanced programming and webpage making methods, some of these web pages just look like the real ecommerce websites. The phishing websites have little visual difference to hoax the victims. Negligent cyber citizens could easily fall into the snare and lose their credit card number, payment account, password, or other important private information. The result is that the disturbance of the normal ecommerce sequence and the victims lose heart to the security of ecommerce. Compared with other crime forms, phishing is relatively new Internet crime. More and more phishing web pages have been found in recent years in an accelerative way (Liu et al., 2006). Ecommerce phishing website is a complicated topic to discuss and analyze, because it is a combination and mixture of social and technical problem. At the moment, there is no an effective prescription to cure it. This paper focus on building an effective method that uses fuzzy data mining algorithms to against the attack of ecommerce phishing websites. Data mining techniques can be used in the fuzzy logic model. It took less time in distinguishing the most important ecommerce phishing website features pointers by analyzing historical data for training purposes.

The paper is organized as follows: Section 2 shows the literature review and related work. Section 3 introduces the theory and methodology for assessing the risk of ecommerce phishing website. Section 4 presents the model design and discussion. Section 5 reveals the conclusion and future work.

2 LITERATURE REVIEW AND RELATED WORK

2.1 Literature Review

Phishing website is only a new form of crime, but it has a destructive influence on ecommerce and other online financial services. Therefore, many parties took steps to prevent and defend such ecommerce phishing website attacks. There are several methods with great promise reported previously. Here we just have a short review of existing anti-phishing solutions and related work. One method is to stop phishing by increasing defense level of email (Adida et al., 2005), because phishing attacks use group mail to tempt victims to click the phishing website (Wu et al., 2006a). The next method is to distinguish

Cheng Y. and Li G..
AN ANTI-PHISHING MODEL FOR ECOMMERCE UNDER A NETWORK ENVIRONMENT.
DOI: 10.5220/0003569404000404
In *Proceedings of the 13th International Conference on Enterprise Information Systems* (NMI-2011), pages 400-404
ISBN: 978-989-8425-53-9

visually the phishing websites from the real ones. Dynamic Security Skins (Dhamija & Tygar, 2005) introduces to adopt a randomly generated visual hash to tailor the interface or web composition to point to the real ones. A common method is to install security toolbars, IE 7 (Sharif, 2006) or higher version web browsers with the phishing filter will block the user's operation when it detect the phishing site. Another method is double factors warranty, which guarantees that the user knows both a secret and a security wand (FDIC, 2004). Nevertheless it has its shortcoming which cannot be authenticated by double factors and cannot protect private information such as credit card information which is not related to a specific website (Wu, et al., 2006b).

Many web browsers use security toolbars to prevent phishing attacks. But some researchers have proven the ineffectiveness through their study. Herzberg and Gbara (2004) introduced a new tool which combines a visual inspection of correct certification with the technique standard certificates. If the certificate was valid, a site-dependent logo would appear in the browser. The credential was provided by a trusted party who has blacklisted existing phishing websites. Bridges and Vaughn (2001), Dhamija and Tygar (2005) presented a scheme that adopts a coded identification and recognition method that allows remote web servers to confirm. While this plan need the whole web base installation, so it is difficult to get supports from the entire industry. In Liu, Deng, Huang, &Fu (2006) proposed a method to model and depict phishing by quantifying a given site's threat, but it is not a solution. Netcraft anti-phishing toolbar (Netcraft, 2004) adopts a blacklist of existing phishing URLs to prevent phishing attacks. The same method also was taken by McAfee, Cloudmark and Microsoft (Pan & Ding, 2006). But the method has its limitation of scalability and timeliness. Given the low-cost short life of phishing websites, APWG proposed an Anti-phishing Working Group (2007) which consists of most of the major anti-phishing companies. Dhamija and Tygar (2005) and Wu (2006b) presented methods with rigid webpage-making rules, random skin or HTML code. But in practice, seldom webpage-makers would take the method (Fu et al., 2006). Then a visual method to detect phishing was first introduced by Liu, Huang, Liu, Zhang, & Deng (2005), Fu et al. (2006), Liu et al., (2006). The method can detect phishing attacks and report in an automatic way. With their efforts, phishing detection begins to focus on visually distinguishable block regions. Fu et al. (2006) clarified the visual similarity in three dimensions: layout, block level and overall style similarity.

2.2 Main Characteristics of Ecommerce Phishing Websites

At present, the crime group consists of professionals with phishing techniques. The phishing websites are highly realistic with sophisticated and undetectable design skills. The criminals always make a phishing website which just looks like a real ecommerce website. They dress up as the goods or service providers with lower price to tempt victims. In fact, there are a plenty of factors and characteristics that can discriminate the real ecommerce websites from the ecommerce phishing ones, such as too long URL address. The detail list is shown in Table 1.

Table 1: Levels and elements of ecommerce phishing websites standard.

Standard	N	Element	Level No.
Domain & URL Identity (Weight=0.4)	1	Abnormal URL	Level 1
	2	Abnormal RUL of Anchor	
	3	Abnormal request URL	
	4	Using the IP address	
	5	Abnormal DNS record	
Java Script & Source Code (Weight=0.2)	1	Server Form Handler	Level 2
	2	Pharming Attack	
	3	Redirect Pages	
	4	Using onMouseOver to hide the Link	
	5	Straddling Attack	
Encryption & Security (Weight=0.1)	1	Abnormal Cookie	
	2	Using SSL Certificate	
	3	Distinguished Names Certificate	
	4	Certification Authority	
Network Address Bar (Weight=0.1)	1	Using @ to confound	Level 3
	2	Displacing analogical chars for URL	
	3	Long URL Address	
	4	Adding a suffix or prefix	
	5	Using Hexadecimal Codes	
Contents & Stylesheets (Weight=0.1)	1	Using Pop-Ups Windows	
	2	Spelling Errors	
	3	Disabling Right-Click	
	4	Copying Websites	
	5	Using Forms with "Submit" Button	
Standard	N	Element	Level No.
Social Human Factors	1	Social General Appellation	
	2	Buying Time to Access Account	
	3	Much stress on Security and Response	
Total Weight			1

3 THE METHODOLOGY TO ASSESS RISK OF ECOMMERCE PHISHING WEBSITES

3.1 The Method of Fuzzy Data Mining

The method of fuzzy data mining is to assess risk of ecommerce phishing website from 27 characteristics from Table 1 which label the risky website. It is useful to represent key phishing characteristics by linguistic variables.

3.1.1 Fuzzifierion

First, we descript every key phishing characteristic indicator by using a range of linguistic descriptors, for example, high, low or medium and we assign different values which from low to high could be divided into classes. In other words, cluster is the degree of belongingness of the values to the membership. An equation is established for every phishing characteristic indicators which shows the mapped input space between 0 and 1. While in decrypting the risk of ecommerce phishing website, linguistic values are assigned as very legal, legal, doubtful, dangerous, risky, very risky. The values of input are between 0 and 10, while for output, range from 0 to 100.

3.1.2 Rule Generation Algorithms

According to the specific risk of ecommerce phishing websites and its key factors, the second step is to clarify the probability varies. The existing rules based on the experts' experience and knowledge in the form of "if⋯then" are useful to assess the probability. We use data mining method and correlation rules to classify the key factors and elements in our ecommerce phishing website model as shown in Fig. 1. There also have some existing research findings about data mining such as Prism (Cendrowska, 1987), C4.5 (Quinlan, 1996), CBA packages (Liu et al., 1998) algorithms to study the relationship of the different phishing elements. All methods above mentioned are easily understood by human (Ciesielski and Lalani, 2003).

3.1.3 Cluster of the Rule Outputs and Defuzzification

Through this step, we can unify the outputs and with the equation of membership to build single fuzzy sets.

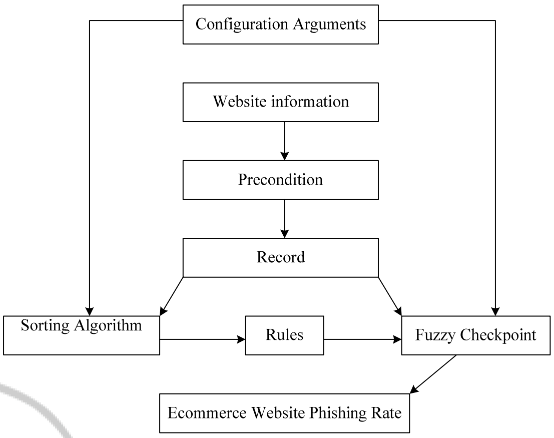

Figure 1: Ecommerce phishing website model.

In this step, we translate a fuzzy output into a crisp output. In the course of fuzziness, we scale the rules, but the end output has to be a number. Centroid technique (Han & Karypis, 2000) was adopted. The output is ecommerce phishing website risk rate and the fuzzy sets from very legal to very risky and defuzzified to get a value.

3.2 Challenges and Classification Algorithms

There are a plenty of challenges to classify ecommerce phishing website. First, it is difficult to collect data of ecommerce phishing websites due to their elusiveness. Secondly, it is a big problem of data life cycle because of ecommerce phishing websites' transience. As estimated, the average phishing site stays live for about 2.25 days (Putting an end to account-hijacking identity theft, 2004). Moreover, it is really difficult to record the original ecommerce phishing website without a mistake. Therefore, it is impossible to find a high accurate algorithm.

In practice, five data mining algorithms such as CBA, Prism, Part, C4.5 and Ripper are very common. They can be used in learning rules from data sets according to different conditions (Misch, 2006).

4 MODEL DESIGN AND DISCUSSION

In this paper, as Table 1 shown, there are six main standards and different elements for every standard. There are 27 elements altogether. There also are three levels on the ecommerce phishing website model as shown in Fig 2. Through classification, cluster and correlation algorithms, we can mine ecommerce phishing website database.

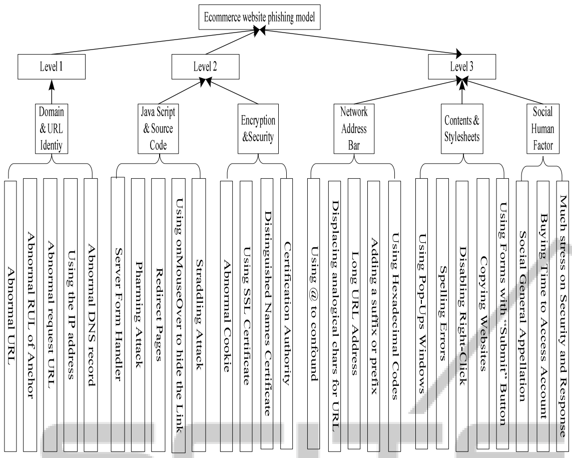

Figure 2: Structure of the model to assess ecommerce phishing website risk rate.

Ecommerce phishing website rating=0.4*Domain and URL Identity crisp [level 1] + ((0.2*Java Script crisp) + (0.1*Encryption crisp &Security crisp)) [level 2] + ((0.1*Contents & Stylesheets crisp) + (0.1* Network Address Bar crisp) + (0.1*Social Human Factor crisp)) [Level 3]

4.1 Rules of Design

4.1.1 Rules for Level 1

In Level 1, there are five input parameters and one output. Every standard has five elements while each element is assumed to be one of three values. While the only one output is the ecommerce phishing website rate fuzzy sets (Fake, Uncertain, Legal) to stand for Domain & URL Identity standard which combines with "If…then" rules to form Rule base 1.

4.1.2 Rules for Level 2

In Level 2, there are two inputs and one output. Through Rule base 1, Java &Source Code standard has five elements while Encryption & Security standard has four elements. The model structure of Level 2 is combined by the two ecommerce phishing website standards, which produces Rule base 2. There is one output rule of the ecommerce phishing website rate fuzzy sets (Fraud, Doubtful or Genuine) indicating Level 2 standard phishing risk rate.

4.1.3 Rules for Level 3

In Level 3, there are three inputs and one output. Through Rule base 1, Contents & Stylesheets standard has five elements. Network Address Bar standard has five elements too. Social Human Factor standard only has three elements. The model structure for Level 3 is combined by the three standards, which produces Rule base 2. There is one output rule of the ecommerce phishing website rate fuzzy sets (Fraud, Doubtful or Genuine) indicating Level 3 standard phishing risk rate.

4.1.4 Rules for End Ecommerce Phishing Website Rate

At last step, there are three inputs, which are respectively Level 1, Level 2 and Level 3. There is one output which is the ecommerce phishing website rate. The model structure is combined by Level 1, Level 2 and Level 3, which produces end ecommerce phishing website rule base. There is one end output fuzzy sets (Very legal, Legal, Doubtful, risky, very risky) indicating end ecommerce phishing website rate.

4.2 Discussion

Because we have no existing data about ecommerce phishing data, we cannot do data mining exercise. We build this model to start the research on ecommerce phishing risk. In this step, some approaches could be adopted. For example, clipping method (Ho, Ling, & Reiss, 2006) could cluster the consequences. Mamdani method (Liu, Chen, & Wu, 2002) could be used to defuzzify the rule evaluation. We believe the 27 characteristics cover the ecommerce phishing website. In the future, if we cooperate with other institutes to collect the recorded data, the validity of the ecommerce anti-phishing model can be tested.

5 CONCLUSIONS AND FUTURE WORK

The ecommerce anti-phishing website model showed the six main standards of the ecommerce phishing website and related twenty seven elements under three levels.

The first purpose in the paper was whether we could find any effective method to detect ecommerce phishing risk automatically. But as mentioned in the paper, it is very difficult to record the data by single researcher. In other words, it requires some other groups’ cooperation to collect the data. In this area, there is more new work to be done and other new techniques should be integrated. However, we believe this model is a good starting for other users to further their research in ecommerce anti-phishing.

REFERENCES

Adida, B., Hohenberger, S. & Rivest, R., (2005). Fighting Phishing Attacks:A Lightweight Trust Architecture fro Detecting Spoofed Emails. In *DIMACS Wkshp on Theft in E-Commerce*.

Anti-Phishing Working Group (2007). Phishing Activity Trends Report. Available from www.antiphishing.org/ reports/apwg_report_april_2007.pdf

Bridges, S. M. & Vaughn, R. B., (2001). Fuzzy data mining and genetic algorithms applied to intrusion detection. *Department of Computer Science Mississippi State University*, White Paper.

Cendrowska, J., (1987). PRISM: An algorithm for inducing modular rules. *International Journal of Man – Machine Studies*, 27(4), 349 – 370.

Ciesielski Vic & Lalani Anand (In press). Data mining of web access logs from an academic web site. In *Proceedings of the third international conference on hybrid intelligent systems (HIS ’ 03)*: Design and Application of Hybrid Intelligent Systems (pp. 1034 – 1043). IOS Press.

Dhamija, R. & Tygar, J. D., (2005). The battle against phishing: Dynamic security skins. In *Proceeding s of the 2005 symposium on Usable Privacy and Security*.

Han, E. & Karypis, G., (2000). Centroid-based document classification: Analysis and experimental results. *Principles of Data Mining and Knowledge Discovery,* 424 – 431.

FDIC (2004). Putting an end to account – hijacking identity theft. Available from http://www.fdic.gov/ consumers/consumer/idtheftstudy/identity_theft.pdf

Herzberg, A. & Gbara, A., (2004). TrustBar: Protecting (even naive) web users from spoofing and phishing attacks. Draft of July, 11.

Ho, C. Y., Ling, B. W. & Reiss, J. D., (2006). Fuzzy impulsive control of high-order interpolative low-pass sigma – delta modulators. *IEEE Transactions on Circuits and Systems* – I: Regular Papers, 53(10).

James, L., (2006). Phishing exposed. Tech target article sponsored by: Sunbelt software.

J. R. Quinlan (1996). Improved use of continuous attributes in c4.5. *Journal of Artificial Intelligence Research*, 4, 77 – 90.

Liu, B., Hsu, W. & Ma, Y., (1998). Integrating classification and association rule mining. In *Proceedings of the fourth international conference on knowledge discovery and data mining. KDD-98, Plenary Presentation*. New York, USA

Liu, M., Chen, D. & Wu, C., (2002). The continuity of Mamdani method. *International Conference on Machine Learning and Cybernetics*, 3, 1680 – 1682.

Liu, W., Deng, X., Fu, A. Y., (2006). Detecting phishing web pages with visual similarity assessment based on earth mover’s distance (EMD). *IEEE Internet Computing*, 3(4)(pp. 58-65).

Liu, W., Deng, X., Huang, G. & Fu, A. Y., (2006). An anti-phishing strategy based on visual similarity assessment, published by the *IEEE computer society (pp. 58-65). Internet Computing IEEE*.

Liu, W., Huang, G., Liu, X., Zhang, M. & Deng, X., (2005). Phishing Web Page Detection. In *Proceeding of eighth International conference on documents analysis and recognition* (pp. 560 – 564).

Misch, S., (2006). Content negotiation in Internet mail. Diploma thesis. *University of Applied Sciences Cologne*, Mat. No.: 7042524.

Netcraft (2004). Available from http://toolbar.netcraft.com.

Olsen, S., (2004). AOL tests caller ID for e-mail.

Pan, Y. & Ding, X., (2006). Anomaly based web phishing page detection. In *Proceedings of the 22nd annual computer security applications conference*.

Wu, M., Miller, R. C. & Garfinkel, S. L., (2006a). Do Security Toolbars Actually Prevent Phishing Attacks? In *Proceedings of the SIGCHI Conference on Human Factors in Computing System*.

Wu, M., Miller, R. C. & Little, G., (2006b). Web wallet: Preventing phishing attacks by revealing user intentions. In *Proceeding s of the Symposium on Usable Privacy and Security*.