

HumanBoost: Utilization of Users' Past Trust Decision for Identifying Fraudulent Websites

Daisuke Miyamoto¹, Hiroaki Hazeyama², and Youki Kadobayashi²

¹ National Institute of Information and Communications Technology
Traceable Network Group
4-2-2 Nukui-Kitamachi, Koganei, Tokyo 184-8795, Japan
daisu-mi@nict.go.jp

² Nara Institute of Science and Technology
Graduate School of Information Science
Internet Engineering Laboratory
8916-5 Takayama,
Ikoma, Nara 630-0192, Japan
hiroa-ha@is.naist.jp,
youki-k@is.naist.jp

Abstract. In this paper, we present an approach that aims to study users' past trust decisions (PTDs) for improving the accuracy of detecting phishing sites. Generally, Web users required to make trust decisions whenever their personal information is asked for by websites. We assume that the database of users' PTDs would be transformed into a binary vector, representing phishing or not, and the binary vector can be used for detecting phishing sites similar to the existing heuristics. For our pilot study, we invited 10 participants and performed a subject experiment in November 2007. The participants browsed 14 emulated phishing sites and 6 legitimate sites, and checked whether the site appeared to be a phishing site or not. By utilizing participants' trust decision as a new heuristic, we let AdaBoost incorporate the heuristic into 8 existing heuristics. The results show that the average error rate in the case of HumanBoost is 9.5%, whereas that in the case of participants is 19.0% and that in the case of AdaBoost is 20.0%. Thus, we conclude that HumanBoost has a potential to improve the detection accuracy for each Web user.

1 Introduction

Phishing is a form of identity theft in which the targets are users rather than computer systems. A phishing attacker attracts victims to a spoofed website, a so-called "phishing site", and attempts to persuade them to provide their personal information. The damage suffered from phishing is increasing. In 2005, the Gartner Survey reported that 1.2 million consumers lost 929 million dollars as a result of phishing attacks [1]. The modern survey conducted in 2007 also reported that 3.6 million consumers lost 3 billion dollars [2].

To deal with phishing attacks, a heuristics-based detection method [3] has begun to garner attention. A heuristic is an algorithm to distinguish phishing sites from the others based on users' experience, and a heuristic checks if a site appears to be a phishing site. For example, checking the life time of the issued website is one of the famous heuristics; the most of the phishing sites' URL are expired in short time period [4]. On the basis of the detection result from each heuristic, the heuristic-based solution calculates the likelihood of a site being a phishing site and compares the likelihood with the defined discrimination threshold. Unfortunately, the detection accuracy of existing heuristic-based solutions is far from suitable for practical use [5], even if various studies discovered heuristics.

In our previous work [6], we attempted to improve the accuracy by employing machine learning techniques for combining heuristics, since we assumed the inaccuracy is caused by heuristics-based solutions that can not use these heuristics appropriately. We evaluated the performance of machine learning-based detection methods (MLBDMs). MLBDMs must achieve high detection accuracy, and they must adjust their detection strategies for Web users, so that the performance metrics consisted of f_1 measure, error rate, and Area Under the ROC Curve (AUC). Our evaluation results showed that the highest f_1 measure was 0.8581, the lowest error rate was 14.15%, and the highest AUC was 0.9342, all of which were observed in the case of the AdaBoost-based detection method. For the most of the cases, MLBDMs performed better than the existing detection method.

In this paper, we propose "HumanBoost", which aims improving the AdaBoost-based detection methods. The key idea of HumanBoost is utilizing Web users' past trust decisions (PTDs). Basically, human beings have the potential to identify phishing sites, even if the existing heuristics cannot detect them. If we could construct the database of PTD for each Web user, it would be able to use the record of the user's trust decision as one feature vector on detecting phishing sites. HumanBoost also involves the idea of adjusting the detection for each Web user. If a user is a security expert, the most predominant factor on detecting phishing sites would be his/her trust decision. Conversely, the existing heuristic will have a strong effect on detection when the user is a novice and his/her PTD has often failed.

As our pilot study, we invited 10 participants and performed a subject experiment. The participants browsed 14 emulated phishing sites and 6 legitimate sites, and checked whether the site appeared to be a phishing site or not. By utilizing participants' trust decision as a new heuristic, we let AdaBoost incorporate the heuristic into 8 existing heuristics. The results show that the average error rate in the case of HumanBoost was 9.5%, whereas that in the case of participants was 19.0% and that in the case of AdaBoost was 20.0%.

The rest of this paper is organized as follows. In Section 2, we summarize the related work, and explain our proposal in Section 3. In Section 4, we describe our

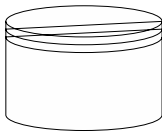
preliminary evaluation, and discuss the limitation of HumanBoost in Section 5. Finally, we conclude our contribution in Section 6.

2 Related Work

For mitigating phishing attacks, machine learning, which facilitates the development of algorithms or techniques by enabling computer systems to learn, has begun to garner attention. PFILTER, which was proposed by Fette et al. [7], employed SVM to distinguish phishing emails from other emails. Abu-Nimeh et al. compared the predictive accuracy of 6 machine learning methods [8]. They analyzed 1,117 phishing emails and 1,718 legitimate emails with 43 features for distinguishing phishing emails. Their research showed that the lowest error rate was 7.72% in the case of Random Forests. Ram Basnet et al. performed an evaluation of 6 different machine learning-based detection methods [9]. They analyzed 973 phishing emails and 3,027 legitimate emails with 12 features, and showed that the lowest error rate was 2.01%. The experimental conditions were different between [8] and [9]; however, machine learning provided high accuracy for the detection of phishing emails.

Apart from phishing emails, machine learning was also used to detect phishing sites. Pan et al. presented an SVM-based page classifier for detecting those websites [10]. They analyzed 279 phishing sites and 100 legitimate sites with 8 features, and the results showed that the average error rate was 16%.

Our previous work employed 9 machine learning techniques [6]. We employed 8 heuristics presented in [11] and analyzed 3,000 URLs, which consisted of 1,500 legitimate sites and the same number of phishing sites, reported on PhishTank.com from November, 2007 to February, 2008. Our evaluation results showed that the highest f_1 measure was 0.8581, the lowest error rate was 14.15%, and the highest AUC was 0.9342, all of which were observed in the case of the AdaBoost-based detection method. In the most of the cases, MLBDMs performed better than the existing detection method.



PTD database

URL	Actual Condition	The user's trust decision	Heuristics #1	..	Heuristics #N
Site 1	Phishing	Phishing	Phishing	..	Legitimate
Site 2	Phishing	Legitimate	Phishing	..	Phishing
Site 3	Phishing	Phishing	Legitimate	..	Phishing
..	..	-
Site M	Legitimate	Legitimate	Legitimate	..	Phishing

Fig. 1. PTD database

3 HumanBoost

The key idea of HumanBoost is utilizing Web users' past trust decisions (PTDs). Generally, Web users are required to make trust decisions whenever they input their personal information into websites. In other words, we assumed that a Web user outputs binary variable, phishing or legitimate, when the website requires users to input their password. It is similar to the existing heuristics.

In HumanBoost, we assume that each Web user has his/her own PTD database, as shown in Fig. 1. The schema of the PTD database consists of the website's URL, actual conditions, the result of the user's trust decision, and the results from existing heuristics. It is to be noted that we do not propose to share the PTD database among users because of privacy concerns.

The PTD database can be regarded as a training dataset that consisted of $N + 1$ binary explanatory variables and 1 binary response variable. Hence, we employ a machine learning technique for studying this binary vector for each user's PTD database. In this study, we employ the AdaBoost algorithm because AdaBoost performed better in our previous work [6].

Further, we expect AdaBoost to cover each user's weak points. Essentially, the boosting algorithms assign high weight to a classifier that correctly labels a site where other classifiers had labeled incorrectly. Assuming that a user's trust decision can be treated as a classifier. AdaBoost would cover users' weak points by assigning high weights on heuristics that can correctly judge the site where he/she is likely to misjudge.

4 Experiment and Results

As our pilot study, in November 2007, we performed a subject experiment by using legitimate enterprise websites and emulated phishing sites. We invited 10 participants who belonged to Nara Institute of Science and Technology, all of them were male, 3 of them completed their M.Eng degree in the last 5 years, while the remaining were master's degree students.

In Section 4.1, we describe how we constructed the dataset, and we explain the design of our experiments in Section 4.2. Finally, we show the results of our preliminary experiment in Section 4.3.

4.1 Dataset Description

Similar to the typical phishing IQ tests performed by Dhamija et al. [12], we prepared 14 emulated phishing sites and 6 legitimate ones, all of which contained Web input forms on which users could input their personal information such as user ID and password. The conditions of the sites are shown in Table 1. Some phishing sites are derived from actual phishing sites according to a report from Phishtank.com. Other phishing sites are emulated phishing sites that mimic actual websites by using phishing subterfuges [7, 13, 14] to induce participants to input their personal information.

Table 1. Conditions in each site

#	Website	Real / Spoof	Lang	Description
1	Live.com	real	EN	URL (login.live.com)
2	Tokyo-Mitsubishi UFJ	spoof	JP	URL(www-bk-mufg.jp)
3	PayPal	spoof	EN	URL (www.paypal.com.%73%69 ... %6f%6d) (URL encoding abuse)
4	Goldman Sachs	real	EN	URL(webid2.gs.com), SSL
5	Natwest Bank	spoof	EN	URL(online-session-0815.natwest.com.esb6eyond.gz.cn) (Derived from PhishTank.com)
6	Bank of the West	spoof	EN	URL (www.bankofthwest.com)
7	Nanto Bank	real	JP	3rd party URL (www2.answer.or.jp), SSL
8	Bank of America	spoof	EN	URL(bankofamerica.com@index.jsp-login-page.com) (URL scheme abuse)
9	PayPal	spoof	EN	URL (www.paypal.com) but first a letter is a Cyrillic small letter a (U+430) (IDN abuse)
10	Citibank	spoof	EN	URL(IP address)
11	Amazon	spoof	EN	URL (www.importen.se), contains amazon in its path (Derived from PhishTank.com)
12	Xanga	real	EN	URL (www.xanga.com)
13	Morgan Stanley	real	EN	URL (www.morganstanleyclientserv.com), SSL
14	Yahoo	spoof	EN	URL(IP address)
15	U.S.D of Treasury	spoof	EN	URL (www.tarekfayed.com) (Derived from PhishTank.com)
16	Sumitomo Mitsui Card	spoof	JP	URL (www.smc-card.com)
17	eBay	spoof	EN	URL (securiry.ebayonlineregist.com)
18	Citibank	spoof	EN	URL (VeCoN.com) (is pronounced “Shi Tei Ban Ku”, look-alike “CitiBank” in Japanese Letter) (IDN abuse)
19	Apple	real	EN	URL (connect.apple.com), SSL, popup warning by accessing non-SSL content
20	PayPal	spoof	EN	URL (www.paypal.com@verisign-registered.com) (URL scheme abuse)

4.2 Design of Experiment

We used a within-subjects design, where every participant saw every website and judged whether the site was deemed a phishing site or not. In our test, we asked 10 participants to browse the websites freely. We installed Windows XP on each participant’s system with Internet Explorer (IE) version 6.0 as the browser. Other than configuring IE to display IDN, we did not install security softwares and/or anti-phishing toolbars. We also did not prohibit participants to access websites that were not listed in Table 1. Therefore, some participants inputted several terms into search engines and compared the URL of the site with the URLs of those listed in Google’s search result pages.

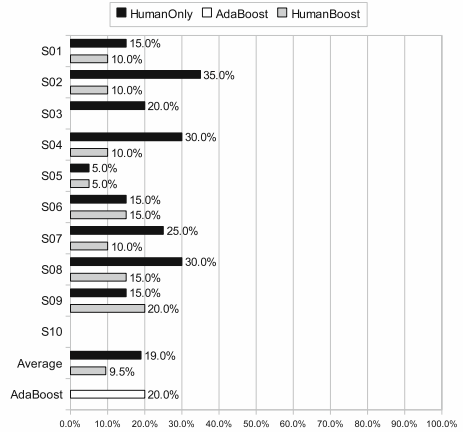
The detection results by each participant are shown in Table 2. In addition, “#” denotes the number of websites in Table 1, P1 - P10 denote the 10 participants, the letter “F” denotes that a participant failed to judge the website, and the empty block denotes that a participant succeeded in judging correctly.

4.3 Results of Experiment

At first, we determined the detection accuracy of the AdaBoost-based detection methods. We employed 8 heuristics, all of which were proposed by Zhang et al. [11] and outputted a binary variable representing phishing or not.

Table 2. The detection result by each participant

#	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
1							F	F		
2	F	F								
3		F								
4			F			F			F	
5							F	F		
6		F	F		F	F				
7	F	F		F		F				
8										
9										
10				F						
11							F			
12							F		F	
13		F	F							
14								F		
15				F						
16		F		F				F		
17		F		F				F		
18										
19	F			F			F	F	F	
20			F							

**Fig. 2.** Error Rate in HumanOnly, AdaBoost and HumanBoost

In this evaluation, we performed 4-fold cross validation to average the result. However, we considered that the experiment involved a small, homogeneous test population, and so it would be difficult to generalize the results to typical phishing victims. We will discuss our plan for removing the bias in Section 5. In addition, we selected 1 as the iteration time, and decided to use the average error rate as a performance metric. On the basis of this condition, we observed that the average error rate of the AdaBoost-based detection method was 20.0%.

Next, we also calculated the detection accuracy of our proposed HumanBoost. We constructed 10 of the PTD database. In other words, we made 10 types of 20×9 binary vectors. Under the same condition described above, we calculated the average error rate for each case of the HumanBoost.

The results were shown in Fig. 2, where HumanOnly denotes a detection accuracy without using AdaBoost and/or HumanBoost. By comparing the case of HumanBoost with the case of HumanOnly, the error rate was lesser or equal in the most of the cases. The average error rate in the case of HumanBoost was 9.5%, whereas the average error rate in the case of HumanOnly was 19.0% and that in the case of AdaBoost was 20.0%. In addition, we performed paired t-test and observed that there was a statistical difference between the accuracy in the case of HumanBoost and that in the case of HumanOnly.

In particular, the average error rate of P9 decreased from 85.0% (HumanOnly) to 80.0% (HumanBoost). We found that some heuristics were assigned higher weights than P9's trust decision. In our experiment, P9 had labeled 3 legitimate sites as phishing sites, whereas the existing heuristics had labeled these 3 sites correctly. Accordingly, the detection of P9 was overwhelmed by that of existing heuristics. We assumed that this is the reason for increasing the error rate.

5 Discussion

Basically, removing bias is important for a participant-based test. Although we used cross validation and paired t-test to eliminate bias, it still can be assumed that there was bias due to the number of samples and/or biased samples. We positioned our laboratory test as a first step, and decided to perform a field test in a large-scale manner. The one approach toward field test is implementing HumanBoost-capable phishing prevention system. It can be possible by distributing the work as browser-extension with some data collection and getting a large population of users to agree to use it.

The weak point of the HumanBoost-capable system is that the system always works after the user finished making the trust decision. Generally, phishing prevention systems will work for users to avoid visiting phishing sites. Apart from these systems, HumanBoost requires users to judge if their secret can be inputted into the site. To protect the users, the HumanBoost-capable system should cancel the input or the submission of users' secret.

Another problem is the difficulty in convincing users to reconsider their trust decision. When a user attempts to browse a phishing site, usual phishing prevention systems display some alert messages that he/she could be visiting a phishing site. In HumanBoost, such messages would be shown after making the trust decision. Otherwise the user recalls his/her trust decision, the HumanBoost-capable system would not block phishing attacks even if the system alerts correctly.

6 Conclusion

In this paper, we presented an approach called HumanBoost to improve the detection accuracy of phishing sites. The key idea was utilizing users' past trust decisions(PTDs). Since Web users might be required of making trust decisions whenever they input their personal information into websites, we considered to record these trust decisions for learning purposes. We simply assumed that the record can be described by a binary variable, representing phishing or not, and found that the record was similar to the output of the existing heuristics.

As our pilot study, we invited 10 participants and performed a subject experiment in November 2007. The participants browsed 14 emulated phishing sites and 6 legitimate sites, and checked whether the site appeared to be a phishing site or not. By utilizing participants' trust decision as a new heuristic, we let AdaBoost incorporate the heuristic into 8 existing heuristics. The results show that the average error rate in the case of HumanBoost was 9.5%, whereas that of participants was 19.0% and that in the case of AdaBoost was 20.0%. Our paired t-test showed that there was a statistical difference. Thus, we concluded that HumanBoost has a potential to improve the detection accuracy for each Web user.

We also discussed to perform our tests in lesser biased ways. To facilitate the field test in a large-scale manner, we mentioned the limitation of HumanBoost from the aspect of the system design. We found that HumanBoost-capable system should have an ability of canceling the input or the submission of users'

secret, instead of blocking phishing sites. We also found that the HumanBoost-capable system should have some interfaces that can expedite users re-making trust decisions.

Apart from its development, we attempt to introduce fuzzy factors for users' trust decision. Users can suspend their trust decision instead of labeling a site as phishing or not. Essentially, machine learning techniques can manipulate quantitative variables, so that we do not adhere to categorical variables in our future work.

References

1. McCall, T., Moss, R.: Gartner Survey Shows Frequent Data Security Lapses and Increased Cyber Attacks Damage Consumer Trust in Online Commerce (2005), http://www.gartner.com/press_releases/asset_129754_11.html
2. McCall, T.: Gartner Survey Shows Phishing Attacks Escalated in 2007; More than \$3 Billion Lost to These Attacks (2007), <http://www.gartner.com/it/page.jsp?id=565125>
3. Chou, N., Ledesma, R., Teraguchi, Y., Boneh, D., Mitchell, J.C.: Client-side defense against web-based identity theft. In: Proceedings of 11th Annual Network and Distributed System Security Symposium (2004)
4. Anti-Phishing Working Group: Phishing Activity Trends Report - Q1 (2008), http://www.apwg.com/reports/apwg_report_Q1_2008.pdf
5. Zhang, Y., Egelman, S., Cranor, L., Hong, J.: Phishing Phish: Evaluating Anti-Phishing Tools. In: Proceedings of the 14th Annual Network and Distributed System Security Symposium (2007)
6. Miyamoto, D., Hazezama, H., Kadobayashi, Y.: An Evaluation of Machine Learning-based Methods for Detection of Phishing Sites. *Australian Journal of Intelligent Information Processing Systems* 10(2), 54–63 (2008)
7. Fette, I., Sadeh, N.M., Tomasic, A.: Learning to detect phishing emails. In: Proceedings of the 16th International Conference on World Wide Web (2007)
8. Abu-Nimeh, S., Nappa, D., Wang, X., Nair, S.: A Comparison of Machine Learning Techniques for Phishing Detection. In: Proceedings of eCrime Researchers Summit (2007)
9. Basnet, R., Mukkamala, S., Sung, A.H.: Detection of Phishing Attacks: A Machine Learning Approach. *Studies in Fuzziness and Soft Computing* 226, 373–383 (2008)
10. Pan, Y., Ding, X.: Anomaly Based Web Phishing Page Detection. In: Proceedings of the 22nd Annual Computer Security Applications Conference on Annual Computer Security Applications Conference (2006)
11. Zhang, Y., Hong, J., Cranor, L.: CANTINA: A Content-Based Approach to Detect Phishing Web Sites. In: Proceedings of the 16th World Wide Web Conference (2007)
12. Dhamija, R., Tygar, J.D., Hearst, M.A.: Why Phishing Works. In: Proceedings of Conference on Human Factors in Computing Systems (2006)
13. Felten, E.W., Balfanz, D., Dean, D., Wallach, D.S.: Web Spoofing: An Internet Con Game. Technical Report 540-96 (Department of Computer Science, Princeton University)
14. Fu, A.Y., Deng, X., Wenyn, L., Little, G.: The methodology and an application to fight against Unicode attacks. In: Proceedings of the 2nd Symposium on Usable Privacy and Security (2006)