Journal of Intelligent Learning Systems and Applications, **2012, 4, 319-329**
http://dx.doi.org/10.4236/jilsa.2012.44033 Published Online November 2012 (http://www.SciRP.org/journal/jilsa)

Behind HumanBoost: Analysis of Users' Trust Decision Patterns for Identifying Fraudulent Websites

Daisuke Miyamoto[1], Hiroaki Hazeyama[2], Youki Kadobayashi[2], Takeshi Takahashi[3]

[1]Information Technology Center, The University of Tokyo, Tokyo, Japan; [2]Internet Engineering Laboratory, Graduate School of Information Science, Nara Advanced Institute of Science and Technology, Nara, Japan; [3]Security Architecture Laboratory, National Institute of Information and Communications Technology, Tokyo, Japan.
Email: daisu-mi@nc.u-tokyo.ac.jp, hiroa-ha@is.naist.jp, youki-k@is.aist-nara.ac.jp, takeshi_takahashi@nict.go.jp

Received May 29th, 2012; revised October 10th, 2012; accepted October 17th, 2012

ABSTRACT

This paper analyzes users' trust decision patterns for detecting phishing sites. Our previous work proposed HumanBoost [1] which improves the accuracy of detecting phishing sites by using users' Past Trust Decisions (PTDs). Web users are generally required to make trust decisions whenever their personal information is requested by a website. HumanBoostassumed that a database of Web user's PTD would be transformed into a binary vector, representing phishing or not-phishing, and the binary vector can be used for detecting phishing sites, similar to the existing heuristics. Here, this paper explores the types of the users whose PTDs are useful by running a subject experiment, where 309 participants browsed 40 websites, judged whether the site appeared to be a phishing site, and described the criterion while assessing the credibility of the site. Based on the result of the experiment, this paper classifies the participants into eight groups by clustering approach and evaluates the detection accuracy for each group. It then clarifies the types of the users who can make suitable trust decisions for HumanBoost.

Keywords: Detection of Phishing Sites; Trust Decision; Credibility of Websites; Machine Learning; Cluster Analysis

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.

To deal with phishing attacks, a heuristics-based detection method has begun to garner attention. A heuristic is an algorithm to identify phishing sites based on users' experience, and checks whether a site appears to be a phishing site or not. Checking the life time of a registered website is well-known heuristic as most phishing sites' Uniform Resource Locator (URL) expires in short time span. Based on 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.

A current challenge of the heuristics-based solutions is improving the detection accuracy. Our proposed HumanBoost [1] aims at improving the machine learning-based detection methods of phishing sites. The key concept of HumanBoost is utilizing Web users' Past Trust Decisions (PTDs), which is the record of users' past decisions. Basically, humans have the potential to identify phishing sites, even if existing heuristics cannot detect them. HumanBoost constructs PTD databases for each Web user, and uses each of the PTD record as a feature vector for detecting phishing sites. For our pilot study, in November 2007, we invited 10 participants and performed a subject experiment. The participants browsed 14 simulated phishing sites and six legitimate sites, and judged whether or not the site appeared to be a phishing site. We utilize participants' trust decisions as a new heuristic and we let Adaptive Boosting (AdaBoost) [2] incorporate it into eight existing heuristics. The results show that the average error rate for HumanBoost was 13.4%, whereas for participants it was 19.0% and for AdaBoost 20.0%.

This paper analyzes the users' trust decision patterns by investigating their decisions making. We assumed that some participants' PTDs were not useful, since we have found such cases that the detection accuracy of HumanBoost was lower than that of AdaBoost. In July 2010, we invited 309 participants to perform a phishing Intelligence Quotient (IQ) test, asked them the reason of that they identified our prepared websites as legitimate or phishing, and analyzed the criterion when they assessed the credibility of the sites. Based on the analysis, we explored useful trust decision patterns, and found that such

people, who were able to utilize their past experience and assessed the credibility by utilizing both URL of the website and security information of the Web browser rather than content of Web page, tended to have useful PTDs for HumanBoost. To the best of our knowledge, this is the first study for orchestrating users' trust decisions and heuristics by machine learning in consideration of their decisions making patterns.

The rest of this paper is organized as follows. Section 2 summarizes the related work, and Section 3 explains our proposal and preliminary evaluation. Section 4 describes our evaluation conditions, and Section 5 shows our experimental results. Section 6 discusses the availability of PTDsand the effectiveness of Extended Validation Secure Socket Layer (SSL) certificates. Finally, section 7 summarizes our contributions.

2. Related Work

This section introduces existing detection method and subject experiment reports as the related work.

2.1. Detection Method for Phishing Sites

There are two distinct approaches for identifying phishing sites. One is URL filtering. It detects phishing sites by comparing the URL of a site where a user visits with a URL blacklist, which is composed of the URLs of phishing sites. Unfortunately, the effectiveness of URL filtering is limited. Sheng *et al*. reported [3] that URL blacklists were ineffective when protecting users initially, as most of them detected less than 20% of phishing sites at hour zero. The rapid increase of phishing sites hinders URL filtering to work sufficiently due to the difficulty of building a perfect blacklist.

Another approach is a heuristic-based method, which can detect phishing sites by calculating the likelihood of being a phishing site. The detection accuracy of existing heuristic-based solutions was, however, far from suitable for practical use. To increase the detection accuracy, Zhang *et al*. developed CANTINA [4], which employeda novel heuristic, named "TF-IDF-Final" heuristic. When the heuristic attempts to identify phishing sites, it feeds the mixture of the domain name of the current website and extracted words from content into Google. If the domain name matches the domain name of the top 30 search results, the website is labeled legitimate.

Aside from developing new heuristics, the combination methods of heuristics were studied. Our previous work [5] employed nine machine learning techniques for detecting phishing sites. By employing eight heuristics presented by CANTINA, we analyzed 3000 URLs, consisting of 1500 legitimate sites and the same number of phishing sites, reported on PhishTank.com [6] from November 2007 to February 2008. Finally, we evaluated the performance of machine learning-based detection methods in comparison to that of CANTINA. Our evaluation results showed the best accuracy wasobserved for the AdaBoost-based detection method. In most cases, machine learning-based detection methods performed better than CANTINA.

2.2. Subject Experiments

Due to the nature of phishing attacks, subject experiments were often used to verify the effectiveness of the countermeasures, such as phishing prevention systems and educational materials against phishing [7,8].

To know how people make their trust decision, Dhamija *et al*. showed 22 participants 20 websites and asked them to determine which ones were fraudulent, and why [9]. They found that 20% of the participants had not looked at the address bar and/or the security indicators, and it led to incorrect choices 40% of the time. Fogg *et al*. observed that 2684 participants evaluated the credibility of two websites and the participants commented about the received signal from the sites [10]. They found that the "design look" of the website was mentioned most frequently, being present in 46.1% of the comments.

3. HumanBoost

This section outlines HumanBoost [1], a mechanism to improve the detection accuracy of phishing sites.

3.1. Overview

The key concept of HumanBoost is utilizing Web users' Past Trust Decisions (PTDs). Web users are generally required to make trust decisions whenever they input their personal information into websites. In other words, we assumed that a Web user outputs a binary variable, phishing or legitimate, when the website requires users to input their password. Note that existing heuristics for detecting phishing sites, all of which were explained in [4], are similar to output binary variables denoting phishing or not phishing.

In HumanBoost, we assume that each Web user has his/her own PTD database. 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. Note that we do not propose sharing the PTD database among users due to the privacy concerns. Given the number of existing heuristics N, the PTD database can be regardedas a training dataset that consists of $N + 1$ binary explanatory variables and one binary response variable. We, therefore, employ a machine learning technique for studying this binary vector for each user's PTD database.

3.2. Theoretical Background

In this study we employ the AdaBoost algorithm that learns a strong algorithm which returns the output H by combining a set of weak algorithms h_t and a set of weight α_t:

$$H = \sum h_t \times \alpha_t \quad (1)$$

The weights are learned through supervised training off-line. Formally, AdaBoost uses a set of input data $\{x_i, y_i : i = 1, \cdots, m\}$ where x_i is the input, y_i is the classification and m is the number of samples.

Each weak algorithm is only required to have an error rate lower than 50%. The AdaBoost algorithm iterates the calculation of a set of weight $D_t(i)$ on the samples. At t = 1, the samples are equally weighted so $D_t(i) = 1/m$.

The update rule consists of three stages. First, AdaBoost chooses the weight α_t as shown in (2).

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) \quad (2)$$

where $\varepsilon_t = \sum_{i=1}^{m} D_t(i)\left[h_t(x_i) \neq y_i\right]$ is the weighted error rate of classifier h_t. Second, AdaBoost updates the weights by (3).

$$D_{t+1} = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{if } h_t(x_i) = y_i \\ e^{\alpha_t} & \text{if } h_t(x_i) \neq y_i \end{cases} \quad (3)$$

where Z_t is a normalization factor, $\sum_{i=1}^{m} D_{t+1}(i) = 1$. Finally, it outputs the final hypothesis Has shown in (1).

In the context of detecting phishing, α_t is the weight for each heuristic h_t, and $D_t(i)$ is the weight for each website while training h_t. If h_t correctly classifies the sites whose $D_t(i)$ is high, ε_t would be low because it only increases when $\left[h_t(x_i) \neq y_i\right]$, hence, α_t would be high as shown in (2). The reason for high $D_t(i)$ is that the other heuristics often fail to label the site (i) correctly. In short, AdaBoost assigns high weight to a classifier that correctly labels a site that other classifiers had labeled incorrectly.

We have two reasons of employing AdaBoost. One is that it had performed better in our previous comparative study [5], where it demonstrated the lowest error rate, the highest f_1 measure, and the highest AUC of the AdaBoost-based detection method, as mentioned in Section 2. The other is that we expect AdaBoost to cover each user's weak points.Assuming that a user's trust decision can be treated as a classifier, AdaBoost would cover users' weak points by assigning high weights to heuristics that can correctly judge a site that the user is likely to misjudge.

3.3. Preliminary Evaluation of HumanBoost

As a pilot study, we invited 10 participants, all Japanese males, from the Nara Institute of Science and Technology. Three had completed their master's degree in engineering within the last five years, and the others were master's degree students. In November 2007, the participants browsed 14 simulated phishing sites and six legitimate sites, as shown in Appendix A. Since all participants lived in Nara Prefecture, we employed the Nanto Bank, a Japanese regional bank in Nara, for website 7.

We used a within-subjects design, where every participant saw every website and judged whether or not it appeared to be a phishing site. In our test we asked 10 participants to freely browse the websites. Each participant's PC was installed with Windows XP and Internet Explorer (IE) version 6.0 as the browser. Other than configuring IE to display International Domain Name (IDN), we installed no security software and/or anti-phishing toolbars. We also did not prohibit participants from accessing websites not listed in Appendix A. Some participants therefore inputted several terms into Google and compared the URL of the site with the URLs of those listed in Google's search results.

By utilizing participants' trust decisions as a new weak hypothesis, we let AdaBoost incorporate the heuristic into eight existing heuristics, namely Age of Domain, Known Images, Suspicious URL, Suspicious Links, IP Address, Dots in URL, Forms, and TF-IDF-Final heuristics all of which were employed by CANTINA [4]. The results show that the average error rate for HumanBoost was 13.4%, whereas that for participants was 19.0% and for the AdaBoost-based detection method 20.0%.

We then conducted a follow-up study. The study had in March 2010, new participants, aged 23 to 30. All were from the Japan Advanced Institute of Science and Technology. All were Japanese males, two had completed their master's degree in engineering within the last five years, and the others were master's degree students. Before conducting the follow-up study, we modified the dataset described in Appendix A. Due to the renewal of PayPal's website during 2007-2010, we updated websites 9 and 20 to mimic the current PayPal login pages. Nanto Bank, website 7 in Appendix A, had changed both the URL and the content of its login page. Nanto Bank is also not well-known in Ishikawa Prefecture, where the participants of the follow-up study lived. We therefore changed website 7 to Hokuriku Bank (another Japanese regional bank in Ishikawa). The domain name of Hokuriku Bank is www2.paweb.answer.or.jp, the same as Nanto Bank.

The invited 11 participants were asked to label 20 websites as legitimate or phishing. Different from the first study, we prepared printed documents to expedite this experiment. Instead of operating a browser, participants looked at 20 screen shots of a browser that had just finished rendering each website. Additionally, showing a

browser screen shot is often used for phishing IQ tests. The results show that the average error rate for HumanBoost was 10.7%, whereas that for participants was 31.4% and for AdaBoost 12.0%.

We found two problems in earlier research. One is that our experiments invited biased sample. All participants were male, and almost of them were belonged the graduate school of information technology, and the rest were received master of engineering. Another is that some participants' PTD were not useful. In the case of such participants, we observed that the average error rate of HumanBoost were higher than that of AdaBoost.

4. Experimental Design

This paper attempts to solve problems described in Section 3.3. In July 2010, we invited various participants to thwart bias and investigated the criterion when participants assessed the credibility of our prepared websites.This section describes how we setupour experiment and the dataset description of the phishing IQ test.

4.1. Experimental Setup

A new phishing IQ test is performed to clarify the people's criterion on judging websites' credibility. We let the participants label these sites as legitimate or phishing, and asked them to the reason of their decisions making by using the form of the questionnaire. The questionnaire items are listed as following.

4.1.1. Past Experience with Website

This is a check of whether a participant has experience of using the websites in Appendix B. In the earlier subject experiment [10], people sometimes drew on their past experiences with a site to reach a conclusion while they were just assessing the credibility of the sites. Accordingly, our participants were asked that they have used the site or not.

4.1.2. Perception of Website's Credibility

This is a check of how a participant labeled a site as legitimate or phishing. Each participant saw options, namely "Content of Web page", "URL of the site", "Security Information of Browser", and "Other Reason". They also marked all that applied (multiple answers allowed), and described their detail reason if selecting "Other Reason" option.

The invited participants saw 20 screen shots of a browser that rendered the websites. These screen shots were taken on Windows Vista and IE 8.0 because IE 6.0 was out of date in July 2010.

After participants finished answering these questionnaires for the websites in Appendix B, we showed the websites in Appendix A. The participants also judged the sites, and we calculated the detection accuracy of each participant, the AdaBoost-based detection method, and HumanBoost.

In the experiment, participants knew that they would be looking at a mixture of phishing sites and legitimate sites. Since we recruited participants via an Internet research company, for a contractual reason we were required to explain the purpose of our experiments to them. We also had to inform them that we never abuse their answer. Hence, the participants realized that they were not personally suffering from our simulated phishing attacks. Observing the activities of the participants in different mental states is interesting, but it is beyond the scope of this paper.

Note that we used IE 8.0 for investigating users' criterion of credibility, and also used IE 6.0 for evaluating the detection accuracy of HumanBoost. The main difference between two conditions was that IE 8.0 was capable of Extended Validation (EV) SSL certificates whereas IE 6.0 was not. With the changes to the SSL certificate interface in modern browsers, a new identity indicator has been introduced to provide a level of confidence in a site's identity. While the site employs an EV SSL certificate, the background of the address bar will be colored green and the information displayed in the area to the right of the lock icon alternates periodically between the organization name/country code, and the CA who issued the certificate. According to the experiments performed by Robert *et al*. [11], EV SSL certificates facilitated for people to identify the ownership of the website. In order to clarify the difference, we planned that both of phishing IQ tests included the website of Tokyo-Tomin Bank which employed an EV SSL certificate. The differences of browser versions will be discussed in Section 6.3.

4.2. Dataset Description

As shown in Appendix A, this paper prepared 14 phishing sites and six legitimate ones. In comparison to the typical phishing IQ test [9] that prepared 13 phishing sites and seven legitimate ones, our existence rate of phishing sites was not so extremely higher.

Our intention of choosing these websites is to show the participants' criterion for credibility. For checking if the participants assess the websites by their past experience, the dataset consisted of websites that the participants were likely to use. We also employed Tokyo-Tomin Bank, which is the one of Japanese regional bank in Tokyo, as website 7 and 25 while we assumed that the participants mainly lived around Tokyo area.

We also let the participants consider when they tried to label correctly. The participants who were likely to check by "URL of the site" would confuse to label the website 26 and 29, since these sites had almost the same URL as the legitimate sites except for one letter. The URLs of the

websites 22, 23, 36, 37 and 39 contained a legitimate-sounding domain name.The websites 25 and 28 were legitimate but the domain name of these sites had no indication of their bland names. For participants who tended to check by "Security Information of Browsers," the websites 29 and 35 might be difficult because they were phishing but equipped with valid SSL certificates. Conversely, websites 34 and 40 were legitimate but did not employ valid SSL certificates though they required users to login. Of course, our prepared phishing websites were look alike of the legitimate ones. It might be difficult if the participant relied on "Content of Web page."

Of the recruited 309 participants, 42.4% (131) were male and 57.6% (178) were female. Age ranged from 16 to 77 years old. 48.2% of participants (149) were office workers, and 19.7% (61) were households and 5.8% (18) were students. Of the students, 66.7% (12) were Bachelors, 11.1% (2) were high school students, 5.6% (1) were masters' degree students.

The other conditions of this study are the same as the follow up study described in Section 3.3. In July 2010, the 309 participants looked at 40 screen shots and judged whether the site seems to be phishing or legitimate.

5. Experimental Results

This section described the result of clustering analysis to explore the types of the participants whose PTDs are useful. After classifying the participants into some groups, we calculate the detection accuracy of each group for comparative study.

5.1. Factors of Ability for Correct Decision

We observed that the error rate was 42.7% if participants answered that they have experience to use the site. Oppositely, that was 48.6% if participants did not. Therefore, we assumed that past experience with the site has positive effectiveness on identifying the site.

Next, we also observed the relationships among the response for the perception of the website's credibility and the error rate, as shown in **Table 1** where a checkmark denotes that the option was selected. Based on the result, we assumed that both labeling by URL of the site and labeling by security information of the browser have positive effectiveness on identifying the site, otherwise labeling by content of Web page has negative effectiveness.

We then analyzed the effectiveness of selecting the "Other Reason" option as to whether our factors should be affected by it. This option was selected by 39 participants, with 170 total times being selected; 2.8% of the all times. According to the participants' descriptions, the main reason for choosing this option was "I don't know"; 20 participants stated it 120 times (2.0%) in total. Of the 120 times, it was three times that the "Other Reason" option was selected with "Content of Web page" option. We regarded that the participant checked the sites by "Content of Web page". The rest of the 117 times, 19 participants selected "Other Reason" option without checking any other options. In our evaluation, these participants were treated as usual as other participants, but in Section 6.2, we discuss the availability of PTD regarding partici- pants who answered "I don't know."

In other cases, we found that the participants checked "Other Reason" instead of checking other options. For example, some participants stated, "I can see the legitimate company's logo" or "The page requires an account number before logging into the website". These answers were regarded as selecting "Content of Web page" option. Such cases were observed for 21 participants in 47 times (0.8%). There were three other cases, but three participants stated that they checked by their experience. As we mentioned that past experience was defined as a factor, accordingly, we decided to ignore the effectiveness of choosing "Other Reason".

Based on these findings, we quantified the users' factors of the ability to make correct decision. Note the ability were derived from past subject experiments [9,10], but unfortunately, we could not find any consensus for their quantification. We calculated the factors as follows by using the available information that we could observe.

5.1.1. Factor 1: Utilization of Past Experience

This is a detection accuracy of the website which a participant has an experience to use. For instance, if the participant had experience of using 10 out of 20 sites and correctly answered 8 of the 10 sites, this variable is 0.8.

5.1.2. Factor 2: Ignoring Signals from Content of Web Page

This is the probability that each participant did not select "Content of Web page" as his/her criteria for credibility. Phishing sites are lookalike legitimate websites; therefore, the content and/or the design of the site give no hints for

Table 1. Perception of website's credibility and error rates.

Content of web page	URL of the site	Security information of browser	The average error rate
v			61.9%
	v		25.5%
		v	36.8%
v	*v*		51.7%
v		*v*	60.0%
	v	*v*	17.9%
v	*v*	*v*	49.8%

judging the site. For instance, if a participant checked "Content of Web page" on six sites, the participants ignored the signals of 14 sites, so this variable is 0.7.

5.1.3. Factor 3: Decision Making by the URL of the Site

This is the probability that each participant selected "URL of the site" as his/her criteria for the credibility. The key difference between phishing sites and legitimate sites is the URL, so the detection based on the difference of the URL would facilitate a correct answer.

5.1.4. Factor 4: Awareness of the SSL Padlock Icon

This is the probability that each participant selected "Security information of browser" as his/her criteria for the credibility when the site showed an SSL certificate. As shown in Appendix B, six out of 20 sites are SSL-enabled. Notice that website 28 and 35 are phishing with valid SSL certificates. In such case, users should check both the SSL padlock icon and URL of the site.

5.1.5. Factor 5: Ignoring Unsuitable Security Information

This is the probability of each participant did not select "Security information of browser" as his/her criteria for the credibility when the site did not employ any SSL certificate. Nevertheless some participant yielded that assessing the site by "Security information of browser" even if they could not identify the SSL padlock icon. The number of the site without an SSL certificate is 14; almost of them were phishing, but website 34 and 40 are legitimate.

5.2. Detection Accuracy of Each Cluster

We then categorized participants into some clusters and explored the characteristics of each cluster. This paper-employed Expectation-Maximization (EM) algorithm [12] as a clustering method.

The number of clusters was eight. While there are no best solutions for the problem of determining the number of clusters to exact, we explored the suitable number based on Bayesian Information Criterion (BIC) [13]. The BIC is the value of the maximized log-likelihood measured with a penalty for the number of parameters in the model.Given conditions explained above, the participants were classified into eight clusters, as shown in **Table 2**, where the first column denotes eight clusters, the second column denotes the number of samples in each cluster, and the rest columns denote a cluster center for each cluster.

The participants in cluster 1 and 4 tended to label a site correctly if they had used the site. They also tended to assess by "URL of the site" rather than "Content of Web page." The difference between cluster 1 and 4 were awareness of the SSL indicators; the participants in cluster 1 checked the "Security information of browsers" carefully, whereas the participants in cluster 4 did not. The participants in cluster 2 tended to rely on "Security information of browsers". However, they might be received incorrect signals from web content as the factor 2 and 5 were lower.

We then evaluated the detection accuracy at each cluster by using the 20 sites listed in Appendix A. Based on the detection results, we calculated the average error rate for each participant group, the AdaBoost-based detection method, and HumanBoost. To perform our evaluation in a less biased way, we employed 4-fold cross validation. Furthermore, our cross validation was repeated 10 times in order to average the result. The results are summarized in **Figure 1**, where the blue bars denote the average error rate of each participant, the red bar denotes the average error rate of the AdaBoost-based detection method, and the yellow bars denote that of HumanBoost.

The average error rate for AdaBoost was 10.5%, for participant was 40.5% and for HumanBoost was 9.6%. The lowest average error rate of the participants group

Table 2. Clustering results of EM algorithm.

Cluster	Participant	Mean of factors				
		1	2	3	4	5
1	48	0.827	0.766	0.787	0.543	0.787
2	30	0.576	0.580	0.479	0.835	0.240
3	61	0.591	0.405	0.674	0.047	0.965
4	17	0.753	0.953	0.953	0.039	0.992
5	71	0.411	0.160	0.168	0.058	0.950
6	11	0.561	0.000	0.014	0.000	1.000
7	34	0.124	0.006	0.008	0.000	1.000
8	37	0.441	0.290	0.177	0.421	0.822

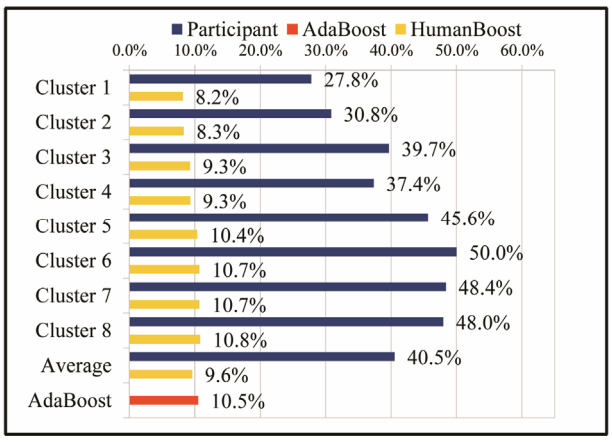

Figure 1. The average error rates of each participant, AdaBoost-based detection method, and HumanBoost.

was 27.8% in cluster 1, followed by cluster 2 (30.8%), 4 (37.4%), 3 (39.7%), 5 (45.6%), 8 (48.0%), 7 (48.4%) and finally 6 (50.0%). The lowest average error rate of HumanBoost was 8.2% in cluster 1, followed by cluster 2 (8.3%), 3 and 4 (9.3%), 5 (10.4%), 6 and 7 (10.7%), and finally 8 (10.8%). In the cases of the participants' cluster 1, 2, 3, 4 and 5, the detection accuracy in HumanBoost was higher than AdaBoost. On the contrary, HumanBoost could not perform better in the cases of the participants' cluster 6, 7, and 8.

From the comparison among each group, we considered that being all of five factorshigher makes HumanBoost performs better. However, we observed that HumanBoost could not utilize the PTDs of the participants in group 6-8, even though their factor 5 is higher. We assumed that the participants in cluster 6 and 7 were novices because they only received signals from "Content of Web page" while assessing the credibility. Due to the lack of knowledge of security and security indicators, they did not selected "Security information of browsers." Even if the cluster 6 has higher value in the factor 1, the performance worse while the other factors were lower. The cluster 8 sometimes checked the security information, but mainly received signals from "Content of Web page" rather than SSL pad-rock icon.

In comparison to the average error rate of HumanBoost in each cluster with that of AdaBoost, the clusters 1 and 2 evidently improve the accuracy by utilizing PTDs. We therefore assumed that a user's PTD might be useful when the user tended to assess the credibility of the URL of the site and/or an SSL indicator of the browser rather than content of Web page.

6. Discussion

This section discusses another approach for investigating the availability of PTD. It then discusses the availability of PTDs of participant who stated "I don't know". It finally explains the effectiveness of differences between the conditions of our phishing IQ tests.

6.1. Availability of PTD from Theoretical Aspect

Theoretically, the key feature of AdaBoost is that a weak hypothesis, which performs just slightly better than random guessing, can be boosted into a strong hypothesis, as we mentioned in Section 3.2. Aside from the clustering approach, it is conceivable that the availability of PTD can be verified by checking the average error rate of each user's PTDs since HumanBoost treats the PTD as one of the weak hypotheses.

We therefore measured the error rate of PTDs for every participant by using 20 websites in Appendix B, classified them into percentile decades, and calculated the average error rate of HumanBoost for each decade by using 20 websites in Appendix A.

Table 3 summarizes the result. We found that the detection accuracy of HumanBoost was still beneficial even if the participant whose PTD has an error ratelower than 50%, but not lower than 60%. The error rates of PTDs in the case of the average error rates of PTDs (= x) was $20\% \le x < 30\%$ was 7.9%, that was lesser than 8.3% in the case of the $10\% \le x < 20\%$.

From these findings, we assumed that participant's criterion for the credibility should be checked whenever verifying the availability of HumanBoost even if the participants could detect websites accurately.

6.2. Participant Who Stated "I Don't Know"

Table 4 summarizes the clustering results of the 19 participants who checked the "Other Reason" option and stated "I don't know" as the reason. In **Table 4**, the first column denotes the eight clusters, the second column denotes the number of the participants who belong to the cluster, and the third column denotes the total times of the site that the participant answered "I don't know".

We then measured the average error rate of the participant sand HumanBoost. The results are summarized in the fourth column and fifth column in **Table 4**. The average error rate of each participant was 45.0%, and that of the

Table 3. The error rate of PTD and HumanBoost.

Error rate of PTD	Number of participant	The average error rate of HumanBoost
$x < 10\%$	4	7.6%
$10\% \le x < 20\%$	16	8.3%
$20\% \le x < 30\%$	40	7.9%
$30\% \le x < 40\%$	36	8.1%
$40\% \le x < 50\%$	37	9.9%
$50\% \le x < 60\%$	63	9.7%
$60\% \le x < 70\%$	56	10.6%
$70\% \le x < 80\%$	54	11.2%
$80\% \le x$	3	12.5%

Table 4. Error rates of participant who stated "I don't know".

Cluster	Participant	Total times	The average error rate	
			PTD	HumanBoost
1	3	10	25.0%	9.7%
2	0	0		
3	2	7	67.5%	7.3%
4	0	0		
5	7	26	52.9%	9.2%
6	0	0		
7	1	1	35.0%	10.5%
8	6	73	46.7%	9.2%

HumanBoost was 9.2%. In comparison to the clustering results in the average error rates for clusters 3, 5, 7 and 8 were lower than with participants who answered "I don't know". However, the average error rate of cluster 1 was higher than that for **Figure 1**.

The results showed that PTDs were usually available even if the participant answered "I don't know." However, we also found that PTDs could not improve the detection accuracy when the participants belonged to cluster 1 who answered "I don't know". We assumed that there were two types of participants who stated "I don't know". One is those who are novices, who have no criteria for identifying the websites. The other is that the participants have their own criterion, but could not identify the site to a lack of knowledge on the websites. In future work, the background of "I don't know" answers.

6.3. Effectiveness of Extended Validation SSL Certificates

Aswe explained in Section 4.1, this paper used IE 8.0 for investigating users' criterion of credibility, and also used IE 6.0 for evaluating the detection accuracy of HumanBoost. In order to clarify the difference, we focused on analyzing the detection accuracy of website 7 in Appendix A and website 25 in Appendix B, these were the legitimate sites of Tokyo-Tomin Bank. Aside from the interfaces for the EV SSL, the conditions were the same in the two sites.

The average error rates for each cluster are shown in **Table 5**. Due to the difficulty of identifying Tokyo-Tomin Bank, the average error rates were tended to be higher; the URL of Tokyo-Tomin Bank started https://www2.paweb.answer.or.jp and the owner of the website was displayed as "NTT DATA CORPORATION", all of which were not associated with Tokyo-Tomin Bank. Albeit such information confused the participants, the average error rate was decreased in the almost all cases of using IE 8.0. Especially, we observed that the detection accuracy of cluster 1, 3, 4 and 8 were dramatically improved. We assumed that the participants regarded the site as phishing by checking URL at first, but reconsidered when they saw the EV SSL certificate. Experiments in the same conditions between the investigation of users' criterion and the evaluation of HumanBoost were our future work.

Table 5. The average error rate of EV SSL capable browser and non EV SSL capable browser.

Cluster	IE 6.0	IE 8.0
1	72.9%	47.9%
2	51.4%	54.1%
3	54.5%	27.3%
4	56.7%	30.0%
5	68.9%	50.8%
6	82.4%	70.6%
7	49.3%	43.7%
8	41.2%	29.4%

7. Conclusions

This study illustrates the trust decision patterns suitable for HumanBoost. For our analysis, we conducted a phishing IQ test with 309 participants in July 2010. First, we investigated the evaluation criteria of websites' credibility from the standpoint of participants. They browsed 14 simulated phishing sites and six legitimate sites, judged whether or not the site appeared to be a phishing site, and answered the reasons of their decision making in the form of questionnaire.

The questionnaire items were "Past Experience with the site," and "Perception of Website's Credibility," the latter required the participants marked all that applied; the options were "Content of Web page", "URL of the site", "Security information of browser" and "Other Reason". Based on their response, we defined five factors of ability for decision making, namely "Utilization of past experience", "Ignoring signals from content of Web page", "Decision making by the URL of the site", "Awareness of the SSL padlock icon" and "Ignoring unsuitable security information."

In order to explore the suitable decision patters for HumanBoost, this study compared the detection accuracy. According to our questionnaire, the participants were classified into eight groups by EM clustering algorithm. The participants also saw another 14 simulated phishing sites and six legitimate sites, judged whether or not the site appeared to be a phishing site. We analyzed that the Past Trust Decisions (PTDs) of the participants who belonged to the particular clusters improved the detection accuracy.

The key finding of our experiments is that the participantswith useful PTDs tend to evaluate sites' URL and/or browser's SSL indicator rather than contents of Web pages to judge the credibility of the sites. This habit leads to make trust decisions correctly. The habit's importance will increase when ordinary Web users start to employ machine learning technique for detecting malicious sites. By integrating the trustful PTDs with state-of-the-art machine learning techniques, we believe the number of phishing incident will be suppressed.

REFERENCES

[1] D. Miyamoto, H. Hazeyama and Y. Kadobayashi, "HumanBoost: Utilization of Users' Past Trust Decision for Identifying Fraudulent Websites," *Journal of Intelligent Learning Systems and Applications*, Vol. 2, No. 4, 2010,

pp.190-199. doi:10.4236/jilsa.2010.24022

[2] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Science*, Vol. 55, No. 1, 1997, pp. 119-139.

[3] S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong and C. Zhang, "An Empirical Analysis of Phishing Blacklists," 2009. http://ceas.cc/2009/main.shtml

[4] Y. Zhang, J. Hong and L. Cranor, "CANTINA: A Content-Based Approach to Detect Phishing Web Sites," *Proceedings of the* 16*th World Wide Web Conference*, Banff, 8-12 May 2007, pp. 649-656. doi:10.1145/1242572.1242659

[5] D. Miyamoto, H. Hazeyama and Y. Kadobayashi, "An Evaluation of Machine Learning-based Methods for Detection of Phishing Sites," *Australian Journal of Intelligent Information Processing Systems*, Vol. 10, No. 2, 2008, pp. 54-63.

[6] OpenDNS, "PhishTank—Join the Fight against Phishing." http://www.phishtank.com.

[7] M. Wu, R. C. Miller and S. L. Garnkel, "Do Security-Toolbars Actually Prevent Phishing Attacks?" *Proceedings of Conference on Human Factors in Computing Systems*, New York, 22-27 April 2006.

[8] P. Kumaraguru, Y. Rhee, A. Acquisti, L. F. Cranor, J. I. Hong and E. Nunge, "Protecting People from Phishing: The Design and Evaluation of an Embedded Training Email System," *Proceedings of Conference on Human Factors in Computing Systems*, San Jose, 27 April-3 May 2007, pp. 905-914.

[9] R. Dhamija, J. D. Tygar and M. A. Hearst, "Why Phishing Works," *Proceedings of Conference on Human Factors in Computing Systems*, New York, 22-27 April 2006, pp. 581-590.

[10] B. J. Fogg, L. Marable, J. Stanford and E. R. Tauber, "How Do People Evaluate a Web Site's Credibility? Results from a Large Study," Technical Report, Stanford, 2002.

[11] R. Biddle, P. C. van Oorschot, A. S. Patrick, J. Sobey and T. Whalen, "Browser Interfaces and Extended Validation ssl Certificates: An Empirical Study," *Proceedings of the 2009 ACM Workshop on Cloud Computing Security*, New York, 9-13 November 2009, pp. 19-30.

[12] A. P. Dempster, N. Laird and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society Series*, Vol. 39, No. 1, 1977, pp. 1-38.

[13] G. E. Schwarz, "Estimating the Dimension of a Model," *Annals of Statistics*, Vol. 6, No. 2, 1978, pp. 461-464. doi:10.1214/aos/1176344136

Appendix A. Conditions of each site used for evaluating the detection accuracy.

#	Website	Legitimate/Phishing	Lang	Description
1	Live.com	Legitimate	EN	URL (login.live.com)
2	Tokyo-Mitsubishi UFJ	Phishing	JP	URL (www-bk-mufg.jp)
3	PayPal	Phishing	EN	URL (www.paypal.com.%73%69 … %6f%6d)(URL Encoding Abuse)
4	Goldman Sachs	Legitimate	EN	URL (webid2.gs.com), SSL
5	Natwest Bank	Phishing	EN	URL (onlinesession-0815.natwest.com.esb6eyond.gz.cn), (Derived from PhishTank.com)
6	Bank of the West	Phishing	EN	URL (www.bankofthevvest.com), similar to the legitimate URL (www.bankofthewest.com)
7	Japanese Regional Bank	Legitimate	JP	URL (www2.paweb.anser.or.jp), SSL (Nanto Bank for the first experiment in November 2007) (Hokuriku Bank for the second experiment in March 2010) (In this study, we employed Tokyo-Tomin Bank)
8	Bank of America	Phishing	EN	URL (bankofamerica.com@index.jsp-login-page.com) (URL Scheme Abuse)
9	PayPal	Phishing	EN	URL (www.paypal.com), first "a'' letter is a Cyrillic small letter "a''(U+430) (IDN Abuse)
10	Citibank	Phishing	EN	URL (IP address)
11	Amazon	Phishing	EN	URL (www.importen.se), contains "amazon'' in its path (Derived from PhishTank.com)
12	Xanga	Legitimate	EN	URL (www.xanga.com)
13	Morgan Stanley	Legitimate	EN	URL (www.morganstanleyclientserv.com), SSL
14	Yahoo	Phishing	EN	URL (IP address)
15	U.S.D. of the Treasury	Phishing	EN	URL (www.tarekfayed.com) (Derived from PhishTank.com)
16	Sumitomo Mitsui Card	Phishing	JP	URL (www.smcb-card.com)
17	eBay	Phishing	EN	URL (secuirty.ebayonlineregist.com)
18	Citibank	Phishing	EN	URL(シテイバンク.com), is pronounced "Shi Tee Ban Ku", look-alike "Citibank" in Japanese Letter)(IDN Abuse)
19	Apple	Legitimate	EN	URL (connect.apple.com), SSL, popup warning by accessing non-SSL content
20	PayPal	Phishing	EN	URL (www.paypal.com@verisign-registered.com), (URL Scheme Abuse)

Appendix B. Conditions of each site used for investigating users' criterion while assessing the credibility of the websites.

#	Website	Phishing/Legitimate	Lang	Description
21	Japan Net Bank	Legitimate	EN	URL(www.japannetbank.co.jp), EV SSL
22	Mizuho MYRAGE Club	Phishing	JP	URL(mizuhobank.biz)
23	mixi	Phishing	EN	URL(mixi-net.net)
24	Yahoo! Japan	Phishing	EN	URL(user-update-09april.com) (Derived from existed phishing domain)
25	Japanese Regional Bank	Legitimate	EN	3rd party URL(www2.answer.or.jp), EV SSL (In this study, we employed Tokyo-Tomin Bank)
26	GungHo Games	Phishing	EN	URL(member.gunho-games.com)
27	Google Mail	Legitimate	JP	URL(www.gmail.com), SSL
28	Mitsubishi-Tokyo UFJ Bank	Legitimate	EN	URL(entry11.bk,mufg.jp), EV SSL
29	Sumitomo Mitsui Card	Phishing	EN	URL(www.smcb-card.com),SSL
30	Twitter	Phishing	EN	URL(capitalmobilehomes.com/?rid=http://twitter.secure.bzpharma.net)
31	Japan Railroad East	Phishing	EN	URL(member.eki-net.com.customer-gdl-7-75.megared.net.mx), (Derived from existed malicious hosts)
32	Amazon	Phishing	EN	URL(ntttokyo0980586.tkyo.ntt.ftth.ppp.infoweb.ne.jp), (Derived from existed malicious hosts)
33	ANA MYRAGE Club	Phishing	EN	URL(IP address)
34	Ameba	Legitimate	EN	URL(www.ameba.jp)
35	Japan Post Holding	Phishing	EN	URL(direct.yucho.org), SSL
36	RAKUTEN	Phishing	JP	URL(rakuten--login.com)
37	SQUARE ENIX	Phishing	EN	URL(secure.playonline-enix.com) (Derived from existed phishing domain)
38	Goo Mail	Phishing	EN	URL(IP address)
39	NICO NICO DOUGA	Phishing	EN	URL(nico-niwango.to)
40	GREE	Legitimate	EN	URL(gree.jp)

