# Users' Behavioral Prediction for Phishing Detection

Lung-Hao Lee[1,2], Kuei-Ching Lee[1,2], Yen-Cheng Juan[1],
Hsin-Hsi Chen[1] and Yuen-Hsien Tseng[2]
[1]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
[2]Information Technology Center, National Taiwan Normal University, Taipei, Taiwan
{d01922005, p00922002, r00922136, hhchen}@ntu.edu.tw, samtseng@ntnu.edu.tw

## ABSTRACT
This study explores the users' web browsing behaviors that confront phishing situations for context-aware phishing detection. We extract discriminative features of each clicked URL, *i.e.*, domain name, bag-of-words, generic Top-Level Domains, IP address, and port number, to develop a linear chain CRF model for users' behavioral prediction. Large-scale experiments show that our method achieves promising performance for predicting the phishing threats of users' next accesses. Error analysis indicates that our model results in a favorably low false positive rate. In practice, our solution is complementary to the existing anti-phishing techniques for cost-effectively blocking phishing threats from users' behavioral perspectives.

## Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: *Information filtering.*

## Keywords
Category prediction, context-aware detection, behavioral analysis.

## 1. INTRODUCTION
Phishing crimes are security threats involving fraudulent web pages that masquerade as trustworthy ones for stealing users' sensitive information, *e.g.*, passwords, personal identification numbers, and credit card numbers. Criminals usually create phishing web pages by exactly copying the legitimate ones or slightly modifying their page content for obtaining users' valuable information. In the past, content-based lexical features have been extracted to detect phishing web sites [3]. A hybrid approach has been proposed to detect phishing web pages by identity discovery and keywords retrieval [2].

Some worldwide users have ever suffered from phishing threats during their web surfing. However, some users never meet phishing dangers. This interesting phenomenon motivates us to study the access contexts in which users will fall into phishing situations from users' behavioral perspectives. Different from previous work that formulates the distinguished patterns between the content of legitimate and phishing web pages, we focus on exploring users' web browsing behaviors to detect phishing threats without crawling web pages for intelligent content analysis.

## 2. USERS' BEHAVIRAL PREDICTION
Users' browsing behaviors on the web result in users' click-through trails, which are defined as access sequences during web surfing. The browsing context of users' information accesses is explored to understand how users fall into phishing threats. The problem statement is described as follows. Let $u_1u_2\ldots u_{(n-1)}u_n$ be a user's access sequence, where $u_i$ is the $i^{th}$ clicked URL in the click-through trail. We focus on determining the category of a user's next access $u_n$, *i.e.*, "Phishing" or "Legitimate", based on the previous accesses $u_1u_2...u_{(n-1)}$ and their contextual information.

We extract 5 discriminative features of each clicked URL in a user's access sequence to capture contextual information for phishing detection. (1) **Domain Name**: phishing URLs tend to look like the original legitimate ones. For example, the domain name "faecbook.com" was verified as a phishing website of social networking service Facebook. We identify the domain names as features for phishing threat detection. (2) **Bag-of-Words**: we first segment a domain name into words delimited by ".". A word is selected if it occurs in a dominant category. Take the domain name www.paypal.com as an example. Only the word "paypal" is retained as a lexical feature, because more than half of its occurrences are rated as the category "Financial Services". In contrast, we can extract a word "paypalsicher" from the domain name www.paypalsicher.eu because this lexical feature always belongs to the category "Phishing". (3) **generic Top-Level Domains (gTLD)**: a URL structure is a hierarchy of names where the upper level consists of a set of Top-Level Domains (TLDs). Security assurance of a URL with gTLD ".gov" (government entities) or ".mil" (military organization) may play more important roles than that with gTLD "info" (informational sites). (4) **IP Address**: phishing criminals usually create and maintain a large number of hosts or redirections to pretend legitimate URLs. These suspected URLs may be hosted in the same suspicious IP address. (5) **Port Number**: Secure Socket Layer (SSL) is a cryptographic protocol that provides communication security on the web. The port number is usually defined as 443 for accomplishing this secure connection. Some phishing pages adopt specific port to achieve their purposes.

We employ the linear chain Conditional Random Field (CRF), which is a type of discriminative probabilistic graph model, by learning users' browsing behaviors for predicting the category of a user's next access. A user's access is regarded as a state in our CRF formulation. Given an observation and its previous states, in terms of the above features, the probability of reaching a state is trained based on Stochastic Gradient Descent. In testing phase, the proposed linear chain CRF reports the category with the largest probability as the result.

## 3. EXPERIMENTS AND EVALUATION
The click-through data, which consists of 76,943 anonymous worldwide users' web browsing behaviors, came from the Trend Micro research laboratory. The category of a user's clicked URL was verified manually from the candidate categories proposed by Trend Micro URL Filtering Engine. Users' click-through trails were divided into two distinct data sets shown as follows. (1) Training set: A phishing trail is denoted as $u_1u_2...u_{(n-1)}u_n$ where the


*WWW'14 Companion*, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577320

previous accesses $u_1u_2...u_{(n-1)}$ are legitimate and the target URL $u_n$ is phishing. Similarly, $u_1u_2...u_{(m-1)}u_m$ represents legitimate trails in which all the accesses are legitimate. Total 99,249 clicked trails from Nov 1st to Dec 31st 2010 were rated as phishing trails. For balanced learning consideration, the same number of legitimate trails was selected. A total of 198,498 users' access trails were used for training. (2) Test set: 134,432 phishing trails from Jan 1st to Mar 15th 2011 were used for testing, and 6,496,860 legitimate access trails from the same time period were used to reflect real-life browsing behaviors.

The following two phishing threat detection approaches based on click-through data were compared to demonstrate their performance. (1) m-gram Hidden Markov Model (**m-gram HMM**): this model adopts category sequences of users' accesses to learn a HMM for security threat prevention [1]. We employ a 4-gram HMM that achieved the best effects for comparisons. (2) Conditional Random Field (**CRF**): this model is the proposed approach for context-aware phishing detection. We also explore different numbers of previous access contexts. This number is denoted as K and is set from 1 to 3 in the experiments.

Table 1 shows the results. The discriminative learning model *CRF* greatly performed better than the generative model *m-gram HMM*. This implies that considering behavioral features extracted from users' access sequences is effective on phishing threat prediction. In addition, the larger the previous access contexts were concerned, the better the precision was achieved. However, shorter contexts accomplished better recall. Considering the tradeoff between precision and recall, the proposed model *CRF* ($K$=2) achieved the best F1 score of 0.9426.

**Table 1. Performance evaluation on phishing detection**

| Models | | Precision | Recall | F1 |
|---|---|---|---|---|
| m-gram HMM | | 0.6735 | 0.5867 | 0.6271 |
| CRF | $K$=1 | 0.9680 | **0.9174** | 0.9420 |
| | $K$=2 | 0.9701 | 0.9167 | **0.9426** |
| | $K$=3 | **0.9702** | 0.9164 | 0.9425 |

We further analyzed the errors of our proposed model *CRF* ($K$=2). Our method maintained a favorably low false positive rate of 0.000586 (*i.e.*, 3807/(6493053 +3807)). We found that most of false positive cases are related to some specific web sites, *e.g.*, the error cases containing the domain name "pr.atwola.com" and "tinyurl.com" were clicked 1,682 and 280 times, respectively. These errors can be avoided with an exception list, which contains legitimate domain names to avoid being incorrectly predicted. We found that some of false negative cases only exist in our test set. This implies that collecting users' access sequences as many as possible is needed for reflecting the characteristics of diversified users' web surfing behaviors, even the big data was analyzed in the experiments. Empirical findings also indicated many users visiting web pages rating as the "Economy," "Shopping," or "Auction" categories, which are all involved in personally financial payments or investments, may fall into phishing situations. It confirms the guideline in which users should be more careful to visit such kinds of web pages for the secured web surfing.

Moreover, we plotted the damage distributions across the countries where phishing-affected victims were located for observing the phishing diffusions. Figure 1 and Figure 2 show the victims' country distribution without and with the help of our approach *CRF* (K=2), respectively. Comparing these two figures, it reveals that our method can effectively decrease severe diffusions of phishing threats. For example, there were 25,124 phishing threats clicked by USA users in the original country distribution. By our prediction model, 76.94% of users can avoid phishing threats. This reflects our model can avoid damages to be propagated unlimitedly from the users' behavioral points of view.

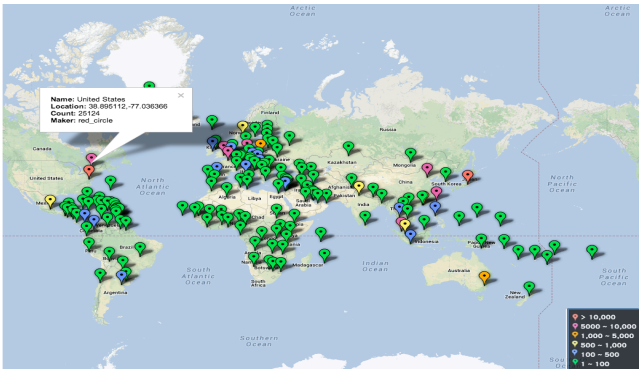


**Figure 1. Country distribution of original phishing-affected victims.**

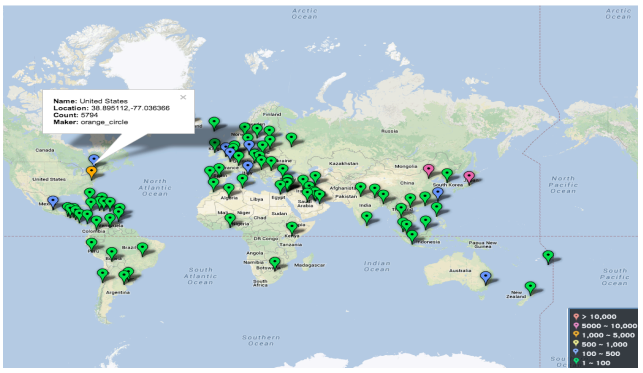


**Figure 2. Country distribution of victims protected by our CRF (K=2).**

## 4. CONCLUSIONS

This work demonstrates the feasibility of exploring users' browsing behaviors only for context-aware phishing detection. Experimental results show that our users' behavioral prediction model, which is complementary to the existing anti-phishing techniques, yields favorable performance on phishing detection.

## 5. ACKNOWLEDGEMENTS

This research was partially supported by National Science Council, Taiwan under grant NSC102-2221-E-002-103-MY3, and the "Aim for the Top University Project" of National Taiwan Normal University, sponsored by the Ministry of Education, Taiwan. We are also grateful to Trend Micro research laboratory for the support of click-through data.

## 6. REFERENCES

[1] Lee, L.-H., Juan, Y.-C., Lee, K.-C., Tseng, W.-L., Chen, H.-H., and Tseng, Y.-H. 2012. Context-aware web security threat prevention. In *Proceedings of CCS*'12. 992-994.

[2] Xiang, G. and Hong, J. 2009. A hybrid phish detection approach by identity discovery and keyword retrieval. In *Proceedings of WWW'09*, 571-580.

[3] Zhang, Y., Hong, J. and Cranor, L. 2007. CANTINA: a content-based approach to detecting phishing web sites. In *Proceedings of WWW'07*, 639-648.