The Novel Features for Phishing Based on User Device Detection

Iuon-Chang Lin[1, 2*], Yi-Lun Chi[3], Hung-Chieh Chuang[1], Min-Shiang Hwang[3]

[1] Department of Computer Science and Engineering, National Chung Hsing University, Taichung, Taiwan.
[2] Department of Photonics and Communication Engineering, Asia University, Taichung, Taiwan.
[3] Department of Computer Science and Information Engineering, Asia University, Taichung, Taiwan.

* Corresponding author. Tel.: +886-4-22840864; email: iclin@nchu.edu.tw
Manuscript submitted February 17, 2015; accepted May 10, 2015.
doi: 10.17706/jcp.11.2.109-115

Abstract: Recent years the rapid developments of technology, Internet services are gradually depending on the environment to be able to provide different services. Due to the rise of mobile devices, in order to provide the most appropriate service to the users, in addition to desktop websites, the most popular sites are beginning to build a websites for mobile devices exclusive service users at the same time. However, phishing website, the phishers will not necessarily build two kinds of websites at the same time.

In this paper, we propose the new phishing features through detect device according the situation. In the experiment, we through the transfer user agent of desktop and mobile to connect, and use SVM to classify, from the result, we find there are the mechanisms in most popular websites, but in phishing websites, the number of this mechanism is rarely.

Key words: Phishing, website, user agent, mobile, desktop.

1. Introduction

With the development of the Internet, people increasingly rely on web applications, web services gradually grow, the value of personal data also rise, and therefore also increase the proportion of personal data against attacks. Phishing websites means phishers imitation of a well-known websites, and deceive the user to obtain the user's confidential information, phishers will provide the URL through variety of social engineering, such as e-mail, to trick the user opens it. If the user clicks the link, as long as the user does not have noticed, the user will think they browsing well-known websites, and according to the site prompted to enter their user accounts and passwords, credit card information, or even personal information.

Once these data sent, phishers can not only get information, but also allow the user unconsciously fall into a trap designed by phisher in the login process, and then leaked their own important personal information. For example, Fig. 1 and Fig. 2, these pages are very similar, if the user does not pay more attention, or use of effective tools to detect phishing, that will not correctly identify the difference between the legitimate and phishing.

Although the relevant research has proposed many detection methods, but still not able to effectively reduce the threat of phishing, phishing sites are still rapidly generated. Phishers in order to lure user fall into the trap, will try a new way to avoid detection current Phishing tools [1]. According to APWG's report which pointed out that the number of Phishing websites in 2014 than in 2013 increased by 10.7 percent. And the number in the fourth quarter of target brand of Phishing attacks increased from 525 in the fourth quarter of 2013 to 557 in the first quarter of 2014 .which payment service is the most important target for

phishing [2].

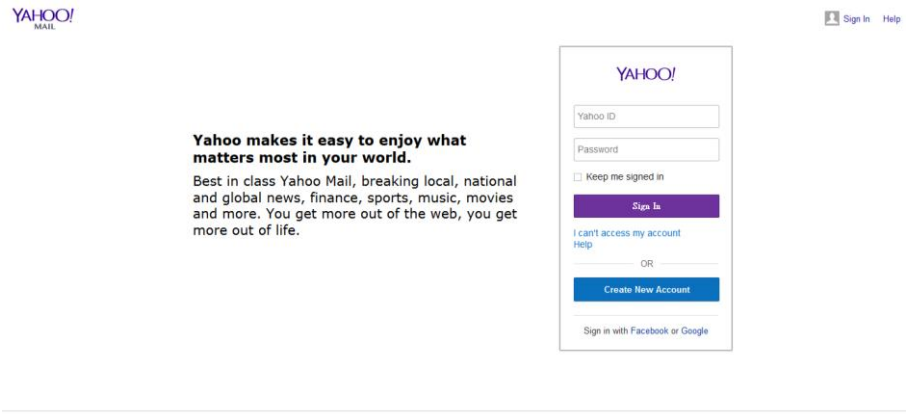

Fig. 1. Yahoo phishing page.

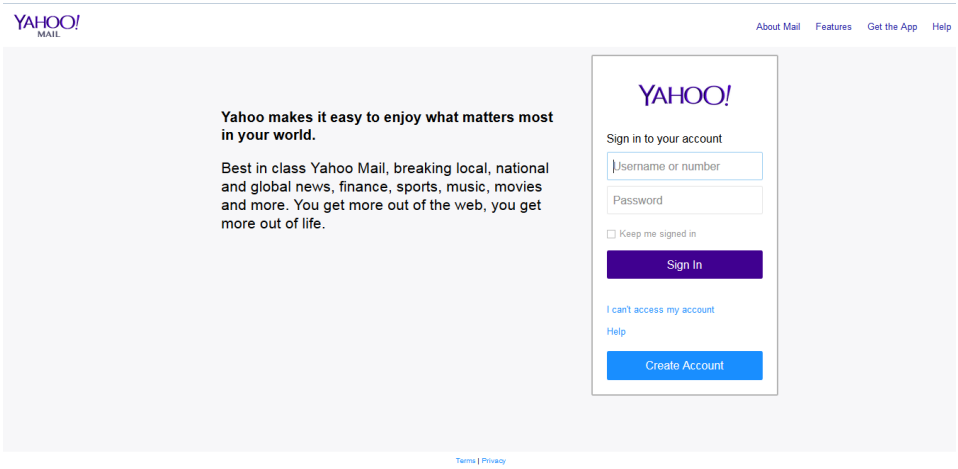

Fig. 2. Yahoo official page.

Recently, due to the development of mobile devices, more and more users will visit the website through mobile device. So, for businesses with an effective mobile website has become particularly important, so well-known companies in order to expand their business development, in addition to the usual desktop websites, will develop mobile websites to give mobile users to browse, such as Fig. 3 and Fig. 4, In addition to corresponding to mobile devices, provides users with a better experience, but also can reduce the load by servers [3], But for phishing websites, in order to fall into the trap to lure user quickly, designing web pages often do not spend too much time, so there may be imitation only a single version of the site, for example, just one desktop website or one mobile site.

So, in this paper, we propose the novel features for detecting phishing websites, and observe the difference between legitimate sites and phishing website.

The rest of this paper, Section 2 will described relevant research on anti-phishing, Section 3 will describe the background and the observation process, Section 4 discusses the relevant observations, the fifth is the paper summation.

2. Related Work

Currently, the method of detecting phishing attacks can be roughly divided into whitelist, blacklist, and heuristics.

Whitelist approach, establishing a white list, if the website does not belong among the list of the site, will

be judged as a suspicious site, but to maintain a valid list is quite difficult, every user's browsing habits are different, is not easy to establish a universal list. Therefore, some studies show that phishing cannot provide the same service to legitimate sites, so gather in login process and use features Naïve Bayesian classifier to determine Phishing websites or not, and record LUI information (Login User Interface) to build their own whitelist user [4]. In [5], it combine whitelist and SVM classifier, When the URL has a high similarity with the whitelist URLs, but it does not the white list, it will be judged as Phishing, if there is the same URL in the list, then recognized as a legitimate site. Low similarity will use an SVM to make predictions based on some features of phishing.

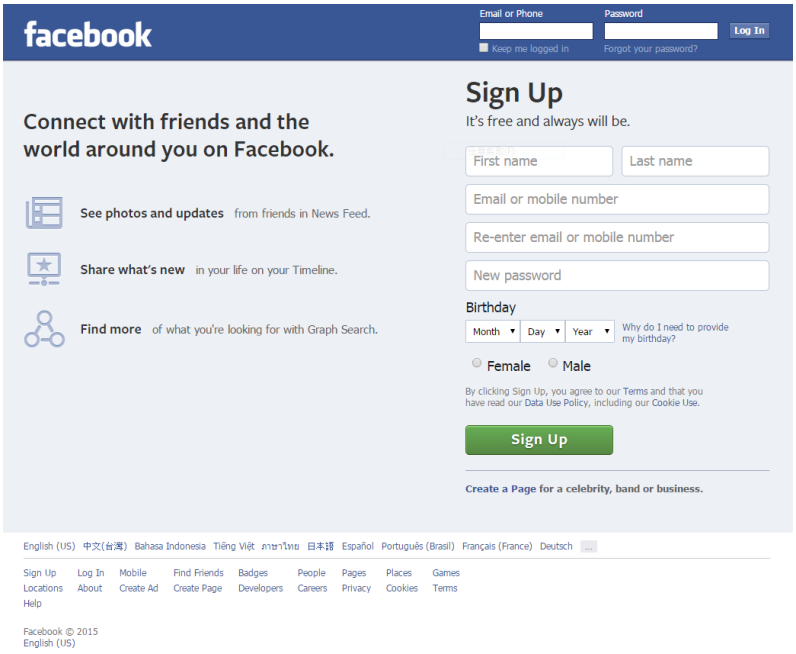

Fig. 3. Facebook desktop page.

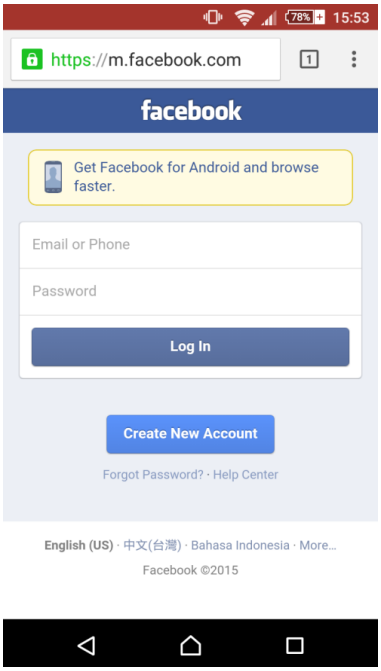

Fig. 4. Facebook mobile page.

Blacklist approach, similar to the white list, however is to create a blacklist of sites when the user browse them in a list, it will prompt user current site is a phishing website, however this way cannot effectively detect some newer phishing website . Therefore, PhishNet [6], Believes that phishers often through simple URL changes, to avoid comparison, first use heuristics of known phishing sites to find new phishing websites, then apply approximate matching algorithm to generate one final score, if the score is greater than threshold, the website will be consider a potential phishing site

Heuristics approach, through the analysis of the site URL and html source code to obtain the relevant features, and with others machine learning methods to determine whether the current page is a phishing site. there is study analyzed 12 features based on based on its content, HTTP transaction, and search engine results [7], and then input to support vector machine classifier, determines whether the webpage is phishing or not. In CANTINA [8], it is a content-based approach to detect phishing, besides the URL and its domain name basically, CANTINA use TF-IDF algorithm to retrieve information, TF-IDF can measure how important a word in a document, and from their experiments result, it is good at detecting phishing, and using TF-IDF can reduce FP rate effectively. In [9], they proposed a method for e-Business websites phishing detection, it combined several features of business websites in Chinese, and use four classification algorithms to decide whether a phishing site. Besides, they consider that researchers should develop domain-specific methods which can detect phishing more useful.

Furthermore, there are some study propose new way different from the past to detect phishing sites,

MobiFish [1] use OCR to detect content among mobile page, turn into text and then compare the current URL, If the URL of the current page does not include the conversion of text through OCR, it will be considered as a phishing site. In 2014, a study [10] through detect the webpage between direct and indirect links, and applying Target Identification(TID) algorithm to detect phishing website and find the target which is mimicked.

3. Background

Some famous sites, in order to provide better convenient services to users, they are provided in two versions to browse, for desktop and mobile devices. Based on this result, we propose a feature to check whether the site contains two versions and observe the results.

Generally, there are two way to build website, one is separate mobile website, another method is Responsive Web Design (RWD), it is a technique that allow you just build a website and that can automatically adapt for different devices ,whether computer or mobile device [11], Business often design a website based on they demand, and the current the number of separate mobile website more than RWD, so we will focus on separate mobile website.

Before loading the site, it will check the user information through the user agent, then decided to load desktop or mobile version site, which will direct users to another domain to load data for example, the desktop website of Facebook is "www.facebook", but mobile website is "m.facebook.com", but some website will retain the original domain and add text express mobile site. Like "www.dropbox.com" and "www.dropbox.com/m".

In addition, some sites do not change in the URL, so we can judge from the number of links. Typically, limited by the performance of mobile devices, and confirmed the brand image, the content of the mobile site will be very simple, easy to mobile users to browse it.

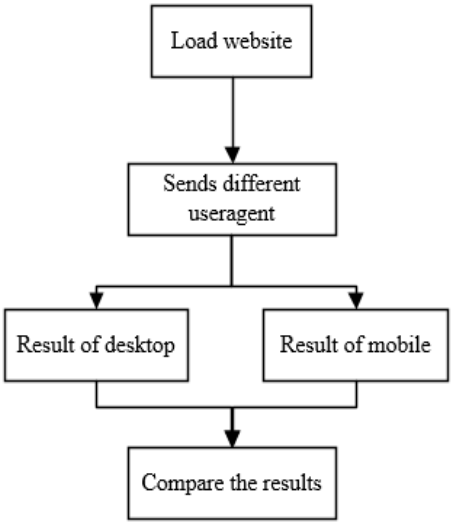

Fig. 5. Flow of detection.

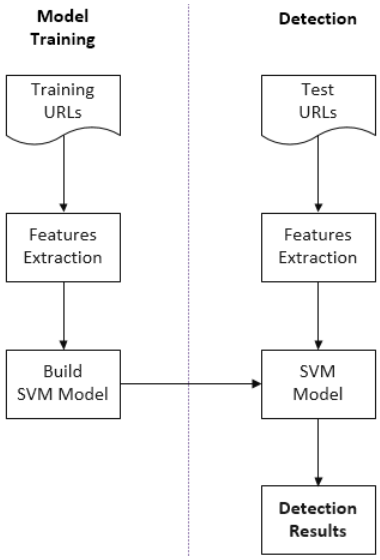

Fig. 6. Flow of Classification.

The main goal of the experiment is to check whether the site contains a mechanism detect user device, and we use the transfer user agent string, to confirm whether the site contains two versions of the website, the desktop version and the mobile version. Fig. 5 describes the entire process of the experiment, first we prepared a desktop and mobile user agent string, through each string to connect, comparing the results after receive data, If the results are the same, the site would be considered that there is no mechanism to detect the user device, In other words, there is only one version. If the result is different, it will think the site has mechanism to detect.

In order to classify the phishing effective, in learning algorithms, we apply SVM (Support Vector Machine) [12] to classify, SVM is machine algorithm which is has been used widely, Because SVM is supervised learning algorithms, we need to divided into training data and test data, in the learning phase, it need to build a model from the training data. Next in test phase, we will determine whether the site is a phishing site or not according the model.

For our implement, we apply LIBSVM [13], a library for Support Vector Machine, it makes we easy to use SVM, and it can solve classification problems effectively.

We divided into two part in the experiment, training and detection, in the training phase, we load the training URLs first. To build the model to classify phishing sites, it generates the relevant features for each page. In the test phase, we load the test URLs, and generate the relevant features for each page. Finally, we apply the model which build in the training phase to determine whether the phishing website. The flow is shown in Fig. 6.

4. Experiments and Results

In this section, we will explain the requirements and sources of the data, and explain the the demand that we proposed the relevant features, and analyzed it. Generally speaking, the phishing attack in order to maximize the benefits, their target always is a well-known site for forged, even a website of bank, but not a website with less traffic, so in the experiment, we analyzed some websites from Alexa which a famous site of analyzed web traffic.

In addition, the phishing sites often have some reason lead they cannot connect, such as they was reported or removed, they only can live a short period of time [14], therefore we tacked some websites which can normal connect in the experiment, it can obtain the relevant features effectively to conduct classification, and get the most reliable results.

We collected the top 100 websites with login service from Alexa [15], and collected 500 phishing can connect at present from Phishtank [16]. In this experiment, we use Java Jsoup to crawl data [17], it provides a simple way to connect and get the links with the website. We generate the 2 different features, one is URL changes, and another is the number of links different.

If the URL changes means it can redirect based on user device, guide users to corresponding website. If the number of links change that means the site provide different content according user device, and there are different number of services.

For classification, we apply LIBSVM [13], and divided into 80% of data for training, 20% for testing, and used two metrics to evaluate. The true positive rate (TPR) and false positive rate (FPR). TPR is the rate of correctly detected phishing in all phishing sites, FPR is the rate of incorrect detected phishing in all legitimate sites.

Table 1. The Result on the Testing Data

	Classified as phishing	Classified as legitimate
phishing	99	1
legitimate	3	17

From Table 1, there are 3 legitimate sites to classify incorrectly, because they use JavaScript which Jsoup is not support, and there are 1 phishing be classified as legitimate, because it was built on others service such as BlogSpot, it can detect by itself. From our experiment, the TPR is 99%, the FPR is 15%.

Few phishing sites with the detect device and redirect users to mobile sites, so in the experiment, the features can distinguish phishing or legitimate successful, our experiment has high TP rate. If the phishing sites were built in other web hosting services, which can detect the devices and set the mobile sites. In such a case, it may cause an error result.

In the legitimate websites, due to the demand of service for mobile devices increased recently, for some websites, they are not build the mobile sites yet, but most businesses have built mobile sites and detect users' devices already.

5. Conclusion

In this experiment, although Jsoup can help us to get data easily, it is not support JavaScript, in the other words, if the site uses other language which Jsoup not support, it will cause to obtain incorrect results.

With the technology development, the services of internet become more diverse, the design methods of build a website also diverse. Relatively, it will more and more approach to bypass old features analysis, increased the threat of phishing. So, in order to reduce the threat effectively, we proposed the novel features, it is according the Internet services at present, the business build a mobile site for mobile users.

In this paper, we present the novel features based on detect device, our experiment result shows that popular legitimate sites will have two versions to provide services. In phishing sites, this situation is not common, so the features are effective for phishing.

In the future work, we can combine with other features, and through other machine learning methods to classify phishing sites to get more accurate results. due to the process of retrieve information and classification both take time and resources, if it combined with other ways to filter phishing sites, for example, it apply blacklist or whitelist to reduce unnecessary detection, it will enhance the accuracy and it can reduce the analysis time effectively.

References

[1] Wu, L., Du, X., & Wu, J. (2014). MobiFish: A lightweight anti-phishing scheme for mobile phones. *Proceedings of IEEE 23rd International Conference on Computer Communication and Networks* (pp. 1-8).

[2] APWG. From: http://docs.apwg.org/reports/apwg_trends_report_q1_2014.pdf

[3] Liu, C. H., & Chen, J. J. Y. (2010). Mobile user agent with user ontology for personalized web service access. *Proceedings of 2010 IEEE International Conference on Systems Man and Cybernetics* (pp. 3956-3960).

[4] Han, W., *et al.* (2012). Using automated individual white-list to protect web digital identities. *Expert Systems with Applications*, *39(15)*, 11861-11869.

[5] Belabed, A., Aimeur, E., & Chikh, A. (2012). A personalized whitelist approach for phishing webpage detection. *Proceedings of IEEE Seventh International Conference on Availability, Reliability and Security* (pp. 249-254).

[6] Prakash, P., *et al.* (2010). PhishNet: Predictive blacklisting to detect phishing attacks. *Proceedings of IEEE INFOCOM* (pp. 1-5).

[7] He, M., *et al*. (2011). An efficient phishing webpage detector. *Expert Systems with Applications*, *38(10)*, 12018-12027.

[8] Zhang, Y., Hong, J. I., & Cranor, L. F. (2007). Cantina: A content-based approach to detecting phishing web sites. *Proceedings of the 16th International Conference on World Wide Web* (pp. 639-648).

[9] Zhang, D., *et al.* (2014). A domain-feature enhanced classification model for the detection of Chinese phishing e-Business websites. *Information & Management*, *51(7)*, 845-853.

[10] Ramesh, G., Krishnamurthi, I., & Kumar, K. S. S. (2014). An efficacious method for detecting phishing webpages through target domain identification. *Decision Support Systems, 61,* 12-22.

[11] Mobile Websites vs Responsive Design: What's the right solution for your business? From: http://adsense.blogspot.tw/2012/07/mobile-websites-vs-responsive-design.html

[12] Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20(3)*, 273-297.
[13] LIBSVM. From: http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[14] Sheng, S., *et al.* (2009). An empirical analysis of phishing blacklists. *Proceedings of Conference on Email and Anti-Spam.*
[15] Alexa. From: http://www.alexa.com/
[16] PhishTank. From: http://www.phishtank.com/
[17] Jsoup. From:http://jsoup.org/

Iuon-Chang Lin received the Ph.D. in computer science and information engineering in March 2004 from National Chung Cheng University, Chiayi, Taiwan. He is currently a professor of the Department of Management Information Systems, National Chung Hsing University, Taichung, Taiwan. His current research interests include electronic commerce, information security, RFID information systems, and cloud computing.

Yi-Lun Chi received her M.S. degrees in management of information systems and technology from School of Information Systems and Technology, Claremont Graduate University, USA in 2006 and in computer science from University of Southern California in 1997. She is currently being an instructor at Overseas Chinese University and pursuing the Ph.D. degree in the Department of Computer Science and Information engineering at Asia University. Her research interests include electronic commerce, internet marketing, data mining, and knowledge management.

Hung-Chieh Chuang was born in 1991. Currently, he is a postgraduate student in the Department of Management Information System, National Chung Hsing University, Taichung, Taiwan. His research interests are in the area of information security.

Min-Shiang Hwang received the B.S. in electronic engineering from National Taipei Institute of Technology, Taipei, Taiwan, Republic of China, in 1980; the M.S. in industrial engineering from National Tsing Hua University, Taiwan, in 1988; and the Ph.D. in computer and information science from National Chiao Tung University, Taiwan, in 1995. He also studied applied mathematics at National Cheng Kung University, Taiwan, from 1984 to 1986. Dr. Hwang passed the National Higher Examination in the field of electronic engineer in 1988. He also passed the National Telecommunication Special Examination the field of information engineering, qualified as an advanced technician in the first class in 1990. From 1988 to 1991, he was the leader of the Computer Center at Telecommunication Laboratories (TL), Ministry of Transportation and Communications, ROC. He was also a project leader for research in computer security at TL in July 1990. He obtained the 1997, 1998, and 1999 Distinguished Research Awards of the National Science Council of the Republic of China. He is a member of IEEE, ACM, and Chinese Information Security Association. His current research interests include database and data security, cryptography, image compression, and mobile communications.