Proc. of the Third Intl. Conf. on Advances in Computing, Electronics and Electrical Technology - CEET 2015

ISBN: 978-1-63248-056-9 doi: 10.15224/ 978-1-63248-056-9-84

Heuristic-based Approach for Phishing Site Detection Using URL Features

Jin-Lee Lee, Dong-Hyun Kim, Chang-Hoon, Lee

***Abstract*—Damage caused by phishing attacks that target personal user information is increasing. Phishing involves sending an email to a user or inducing a phishing page to steal a user's personal information. This type of attack can be detected by blacklist-based detection techniques; however, these methods have some disadvantages and the numbers of victims have therefore continued to increase. In this paper, we propose a heuristic-based phishing detection technique that uses uniform resource locator (URL) features. We identified features that phishing site URLs contain. The proposed method employs those features for phishing detection. The technique was evaluated with a dataset of 3,000 phishing site URLs and 3,000 legitimate site URLs. The results demonstrate that the proposed technique can detect more than 98.23% of phishing sites.**

***Keywords*—phishing sites, URL-based features, heuristic, machine learning**

I. Introduction

With the recent growth of the Internet environment and diversification of available web services, web attacks have increased in quantity and advanced in quality. Phishing is a type of social engineering attack that targets a user's sensitive information through a phony website that appears similar to a legitimate site, or by sending a phishing email [1]. According to research of the Anti Phishing Working Group (APWG), 85,062 phishing sites were globally detected in the second quarter of 2010; by the second quarter of 2014, 128,978 were detected. These figures mark an increase of 1.5 times the value that count of occurred phishing attack in one quarter [2,3]. In addition, annual damage caused by phishing was measured at $5.9 billion. Thus, phishing is a worldwide malicious activity that continues to increase.

In response to this increase in phishing attacks, phishing detection techniques have been the focus of considerable research. Typical phishing detection techniques include the blacklist-based detection method and the heuristic-based technique. The blacklist-based technique maintains a uniform resource locator (URL) list of sites that are classified as phishing sites; if a page requested by a user is present in that list, the connection is blocked [4]. This technique is commonly used and has a low false-positive rate; however, its accuracy is determined by the quality of the list that is maintained. Consequently, it has the disadvantage of being unable to detect temporary phishing sites [5].

Jin-Lee Lee, Dong-Hyun Kim, Changhoon, Lee
Konkuk University
Republic of Korea

** This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea and NAVER Corp., under ICT/SW Creative research program supervised by the NIPA(National IT Industry Promotion Agency) (NIPA-2014-A015-0020)

The heuristic-based detection technique analyzes and extracts phishing site features and detects phishing sites using that information [6]. In this paper, we propose a new heuristic-based phishing detection technique that resolves the limitation of the blacklist-based technique. We implemented the proposed technique and conducted an experimental performance evaluation. The proposed technique extracts features in URLs of user-requested pages and applies those features to determine whether a requested site is a phishing site. This technique can detect phishing sites that cannot be detected by blacklist-based techniques; therefore, it can help reduce damage caused by phishing attacks.

The remainder of this paper is organized as follows. In Section 2, we present related works about phishing sites and phishing detection techniques. The heuristic-based phishing detection technique that employs URL-based features is described in Section 3. In Section 4, we present the evaluation results. In Section 5, we provide our conclusions and describe future work.

II. Related Works

Phishing is an attempt to steal a user's personal information typically through a fraudulent email or website [1]. We conducted a study on phishing sites, which are either fake sites that are designed to appear similar to legitimate sites or sites that simply have phishing-related behaviors. Almost all phishing sites include the functionality in which users enter sensitive information, such as their personal identification, password, and/or account number. These sites can include links to connect to other phishing sites and malicious code that contaminates a user's computer.

Phishing detection techniques can be generally divided into blacklist-based and heuristic-based approaches. The blacklist-based approach maintains a database list of addresses (URLs) of sites that are classified as malicious. If a user requests a site that is included in this list, the connection is blocked [4]. The blacklist-based approach has the advantages of easy implementation and a low false-positive rate; however, it cannot detect phishing sites that are not listed in the database, including temporarily sites [5].

The heuristic-based approach analyzes phishing site features and generates a classifier using those features [6]. When a user requests a web page, the classifier determines whether that page is a phishing site. This approach can detect new phishing sites and temporary phishing sites because it extracts features from the requested web page. Nevertheless, it has the disadvantage of being difficult to implement; moreover, generating a classifier is time-intensive. Thus, the two approaches have both advantages and disadvantages. Therefore, these approaches are selectively employed in the proposed technique depending on the application.

Proc. of the Third Intl. Conf. on Advances in Computing, Electronics and Electrical Technology - CEET 2015

ISBN: 978-1-63248-056-9 doi: 10.15224/ 978-1-63248-056-9-84

III. Proposed Approach

A. URL Structure

A URL is a protocol that is used to indicate the location of data on a network. The URL is composed of the protocol,

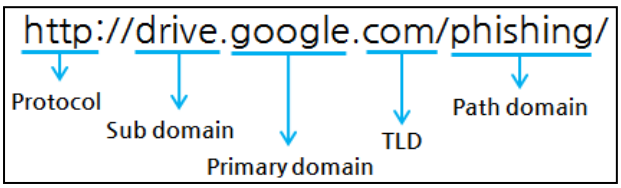

Figure 1. URL Structure.

subdomain, primary domain, top-level domain (TLD), and path domain. [6]. In this study, the subdomain, primary domain, and TLD are collectively referred to as the domain. Fig. 1 depicts the individual components of a URL.

The protocol refers to a communication protocol for exchanging information between information devices; e.g., HTTP, FTP, HTTPS, etc. Protocols are of various types and are used in accordance with the desired communication method.

The subdomain is an ancillary domain given to the domain and has various types depending on the services provided by the domain page. The domain is the name given to the real Internet Protocol (IP) address through the Domain Name System (DNS). The primary domain is the most important part of a domain. The TLD is the domain in the highest position in the domain name hierarchy architecture; e.g., .com, .net, .kr, .jp, etc. [7]. We define features of each component of the URL; these features are used for phishing site detection.

B. URL Features

Table 1 shows 26 URL-based features that are used in the proposed detection technique.

- Features 1 to 6 relate to Google Suggestion. They return a suggested word when a user enters a single term. We analyze the results of Google Suggestion when entering the URLs of phishing sites and legitimate sites. If a search term is similar to a suggested result, input URL is doubtful because that site may be emulating an existing site. We use Levenshtein distance between the two terms—the Google Suggestion result and the search term—as a feature for detecting phishing sites [6,8]. In addition, if a suggested result is the same as that of a domain that is present in the trustworthy whitelist, that search term site may be emulating a legitimate site [8]. For this reason, we can detect phishing sites using this feature.
- Features 7 to 9 can be extracted through page ranking. The page rank is a numerical value that is calculated by the number of visitors and degree of popularity. Phishing sites have a very low page rank value or no value because phishing sites are not often visited by many people and they exist for a short time [9]. Therefore, if a domain page rank value is very low, it can be regarded as a phishing site.
- Features 10 to 14, and 16, 17, 19, and 20, are associated with suspicious URL patterns and characters. Characters such as '@' and '//' rarely appear in a URL. Moreover, URLs of legitimate sites typically have one TLD. Therefore, patterns of many TLDs in a URL signify a fraudulent site [10]. Therefore, in the above cases, we classify these sites as phishing sites.
- Features 21 to 26 are characterized by URL property values. Because temporary phishing sites, as mentioned, often do not contain the required property values [11], property values can be used as features for identifying phishing sites.

TABLE I. URL-BASED FEATURES

No.	URL-based features	
	Feature name	***Description***
1	Similarity of primary domain and Google Suggestion (primary domain)	Levenshtein distance between primary domain and Google Suggestion (primary domain)
2	Similarity of subdomain and Google Suggestion (subdomain)	Levenshtein distance between primary domain and Google Suggestion (subdomain)
3	Similarity of path domain and Google Suggestion (path domain)	Levenshtein distance between primary domain and Google Suggestion (path domain)
4	Safety of Google Suggestion (primary domain)	Whether result of Google Suggestion (primary domain) is present in the whitelist
5	Safety of Google Suggestion (subdomain)	Whether result of Google Suggestion (subdomain) is present in the whitelist
6	Safety of Google Suggestion (path domain)	Whether result of Google Suggestion (path domain) is present in the whitelist
7	Google page rank	PageRank value of domain
8	Alexa rank	AlexaRank value of domain
9	Alexa reputation	Alexa reputation value of domain
10	Via IP address	Whether domain is in the form of an IP address
11	Length of URL	Length of URL
12	Suspicious character	Whether URL has '@', '//'
13	Prefix and suffix	Whether URL has '-'
14	Number of subdomain	Number of dots in domain
15	Length of subdomain	Length of subdomain
16	Port number matching	Whether explicit port number and protocol port number are equal
17	Number of TLD and out of TLD position	More than one TLD in URL, and out of TLD position
18	Phishing words in URL	Whether URL has phishing terms
19	Primary domain spelling mistake	Whether primary domain is similar to whitelisted domains
20	Number of '/'	Number of '/' in URL
21	Country matching	TLD country, and domains country are equal or not
22	HTTPS protocol	Whether URL use https.
23	DNS record	Whether URL has DNS record
24	WHOIS record	Domain age in WHOIS record
25	Value of TTL	TTL value of domain
26	PTR record	Whether domain has PTR record

Thus, many URL features exist that have been employed in several studies on phishing detection. In the present research, we incorporate features used in previous studies

Proc. of the Third Intl. Conf. on Advances in Computing, Electronics and Electrical Technology - CEET 2015

ISBN: 978-1-63248-056-9 doi: 10.15224/ 978-1-63248-056-9-84

and define two new features for identifying existing phishing sites.

- Feature 15 is defined for identifying newly created phishing sites with the proposed technique. Currently, to prevent a user from recognizing that a site is not legitimate, phishing sites typically hide the primary domain; the URLs of these phishing sites have unusually long subdomains. Therefore, we added a feature that calculates the subdomain length of a URL to determine if it is a phishing site. This feature can be additionally used to identify phishing sites that target vulnerabilities of smartphones, which have small displays that make it difficult to see the full URL.
- Feature 18 is another new feature that reflects current phishing trends. This feature includes eight words that are predefined as phishing terms. It is verified if a requested page's URL contains these phishing terms [12][1]. This feature worked well in previous studies; however, we determined that changes have occurred since the studies were conducted. Thus, through experiments, we identified eight new phishing terms[2] and we employ them in our phishing detection technique.

As noted above, our proposed method employs new features that have not been previously used in studies. IN addition, it advances features from previous works to provide better phishing detection performance.

C. *Architecture*

Fig. 2 illustrates the proposed phishing detection process, which includes two phases: training and detection.

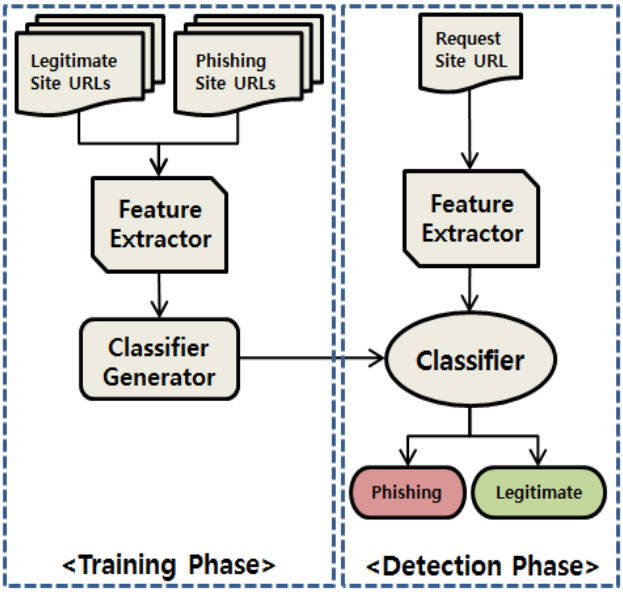

Figure 2. Process of proposed detection method.

In the training phase, a classifier is generated using URLs of phishing sites and legitimate sites collected in advance. The collected URLs are transmitted to the feature extractor, which extracts feature values through the predefined URL-based features. The extracted features are stored as input and passed to the classifier generator, which generates a classifier by using the input features and the machine learning algorithm.

In the detection phase, the classifier determines whether a requested site is a phishing site. When a page request occurs, the URL of the requested site is transmitted to the feature extractor, which extracts the feature values through the predefined URL-based features. Those feature values are inputted to the classifier. The classifier determines whether a new site is a phishing site based on learned information. It then alerts the page-requesting user about the classification result.

D. *Algorithms*

To determine a classifier with the best performance for using URL-based features, we employed several machine learning algorithms: support vector machine (SVM), naive Bayes, decision tree, *k*-nearest neighbor (KNN), random tree, and random forest.

- SVM is a classification method that was introduced in 1992 by Boser, Gyon, and Vapnik [13]. It is a statistical learning algorithm that classifies samples using a subset of the training samples, called support vectors. SVM is built on the structural risk minimization principle for seeking a decision surface that can separate data points into two classes with a minimal margin between them [14]. The advantage of SVM is its capability of learning in spare high-dimensional spaces with very few training samples.
- Decision tree is a classification method that was introduced in 1992 by Quinlan [15]. It creates a tree form for classifying samples. Each internal node of the tree corresponds to a feature, and the edges from the node separate the data based on the value of the feature [15]. Decision tree includes a decision area and leaf node. The decision area checks the condition of the samples and separates them into each leaf node or the next decision area. The decision tree is very fast and easy to implement; however, it has the risk of overfitting.
- Naive Bayes is a classifier that can achieve relatively good performance on classification tasks. It is based on the elementary Bayes' theorem. It is one of most successful learning algorithms for text categorization [16]. On account of the conditional model's feature, naive Bayes is effectively trained in supervised learning. It provides the advantage of learning essential parameters using small training samples.
- KNN is a non-parametric classification algorithm [17]. It has been successfully applied to various information-retrieval problems. It classifies the input data using *k* training data that is similar to the input data. KNN uses Euclidean distance to calculate the similarity between the input and training samples. Its performance is determined by

1) secure, account, websrc, ebayisapi, signin, baking, confirm, login

2) index, includes, content, images, admin, file_doc, paypal, login

Proc. of the Third Intl. Conf. on Advances in Computing, Electronics and Electrical Technology - CEET 2015

ISBN: 978-1-63248-056-9 doi: 10.15224/ 978-1-63248-056-9-84

the choice of *k*; nevertheless, choosing a suitable *k* value is not easy.

- Random tree is a tree-based classification method that was introduced by Breiman and Cutle [18]. A tree is drawn at random from a set of possible trees. "At random" means that each tree in the set of trees has an equal chance of being sampled [18]. The random tree classifier takes as input a features vector and classifies it with each tree. The output is determined by the majority "vote." This algorithm can handle both classification and regression.
- Random forest is a classification method that combines many tree predictors; each tree depends on the values of a random vector that is independently sampled [19]. All trees in the forest have the same distribution. This algorithm can handle a large number of variables in the dataset; however, it lacks reproducibility because the process of forest building is random [20].

IV. Evaluation

To conduct classifier training and evaluation through an experimental dataset, we collected the URLs of phishing and legitimate sites. We gathered 3,000 phishing site URLs from PhishTank and 3,000 legitimate site URLs from DMOZ. The evaluation was conducted using *k*-fold cross validation. *K*-fold cross validation divides the input data into *k*; $k - 1$ datasets are used for training, and the remaining one is used for validation. This process is performed *k* times, such as the number in the divided dataset, because all datasets can be used for training and validation. This method is typically used to evaluate the accuracy of the classifier with a small dataset. In this study, we used ten-fold cross validation to evaluate our detection technique. We performed the testing with the WEKA open-source machine learning tool, and we analyzed the performance of each of the machine learning algorithms noted in Section 3. The accuracy was calculated as TP (true positive), TN (true negative), FP (false positive), and FN (false negative). We compared the performance of each classifier using the calculated accuracy. Fig. 3 depicts the TP, TN, FP, and FN matrix.

		Prediction	
		Positive	Negative
Actual	True	*True Positive*	*False Negative*
	False	*False Positive*	*True Negative*

Figure 3. TP, TN, FP, FN matrix

TP is the ratio of the prediction that a determined phishing site is indeed a phishing site, and FN is the ratio of the prediction that a determined phishing site is actually a legitimate site. In addition, FP is the ratio of the prediction that a truly legitimate site is a phishing site, and TN is the ratio of prediction that a determined legitimate site is indeed a legitimate site. Table 2 shows the TP, TN, FP, FN ratios of each machine learning algorithm.

As a result of the experiments, we obtained TP, TN, FP, and FN ratios to calculate three measurements that we used to compare the performance of each algorithm. The first measurement was for the specificity of the true negative rate. The second was the sensitivity of the true positive rate. The third was the accuracy of the total ratio of the prediction that a determined phishing site is actually a phishing site, and that a determined legitimate site is indeed legitimate.

TABLE II. TF, TN, FP, AND FN OF MACHINE LEARNING ALGORITHMS

Algorithm	Measurements			
	TP	***TN***	***FP***	***FN***
SVM	97.00%	94.90%	5.10%	3.00%
Decision Tree	96.90%	96.90%	3.10%	3.10%
Naive Bayes	90.90%	95.10%	4.90%	9.10%
KNN (k = 1)	96.30%	96.00%	4.00%	3.70%
Random Tree	96.10%	96.00%	4.00%	3.90%
Random Forest	98.10%	98.40%	1.60%	1.90%

In measuring the classifier performance, (1) was the equation of specificity, (2) was the equation of sensitivity, and (3) was the equation of accuracy.

$$Specificity = \frac{TP}{TP+FP} \quad (1)$$

$$Sensitivity = \frac{FP}{FP+FN} \quad (2)$$

$$Accuracy = \frac{TP+TN+FP+FN}{TP+TN} \quad (3)$$

We additionally used the specificity, sensitivity, and accuracy as measurements of the classifier performance measurement. Table 3 shows the specificity, sensitivity, and accuracy of each machine learning algorithm that we used in training.

TABLE III. EXPERIMENTAL RESULTS OF ALGORITHMS

Algorithm	Measurements		
	Specificity	***Sensitivity***	***Accuracy***
SVM	96.93%	95.00%	95.95%
Decision Tree	96.90%	96.90%	96.88%
Naive Bayes	91.26%	94.88%	93.01%
KNN (k = 1)	96.28%	96.01%	96.18%
Random Tree	96.09%	96.00%	96.03%
Random Forest	98.10%	98.30%	98.23%

As a result of the experiment, we determined that the best machine learning algorithm, random forest, used URL features. This classifier detected more than 98.23% of phishing sites. The high accuracy shown in Table 3 and low false-positive rate shown in Table 2 meant that the proposed phishing detection technique can effectively classify sites as either being phishing or legitimate..

V. Conclusion

In this paper, we proposed a heuristic-based phishing detection technique that employs URL-based features. The method combines URL-based features used in previous studies with new features by analyzing phishing site URLs. Additionally, we generated classifiers through several machine learning algorithms and determined that the best

ISBN: 978-1-63248-056-9 doi: 10.15224/ 978-1-63248-056-9-84

classifier was random forest. It showed a high accuracy of 98.23% and a low false-positive rate. The proposed technique can provide security for personal information and reduce damage caused by phishing attacks because it can detect new and temporary phishing sites that evade existing phishing detection techniques, such as the blacklist-based technique.

In future work, we intend to address the time-intensive disadvantage of the heuristic-based technique. With a large number of features, it is time-consuming for the heuristic-based approach to generate classifiers and perform classification. Therefore, we will apply algorithms to reduce the number of features and thereby improve performance. In addition, we will examine a new phishing detection technique that uses not only URL-based features, but also HTML and JavaScript features of web pages to improve performance.

References

[1] Khonji, Mahmoud, Youssef Iraqi, and Andrew Jones. "Phishing detection: a literature survey." Communications Surveys & Tutorials, IEEE 15.4 (2013): 2091-2121.

[2] Anti Phishing Working Group. (2015. March.) APWG Phishing Activity Trend Report 2nd Quarter 2010. [Online]. Available: http://docs.apwg.org/reports/apwg_trends_report_q2_2014.pdf

[3] Anti Phishing Working Group. (2015. March.) APWG Phishing Activity Trend Report 2nd Quarter 2014. [Online]. Available: http://docs.apwg.org/reports/apwg_report_q2_2010.pdf

[4] Huang, Huajun, Junshan Tan, and Lingxi Liu. "Countermeasure techniques for deceptive phishing attack." New Trends in Information and Service Science, 2009. NISS'09. International Conference on. IEEE, 2009.

[5] Ma, Justin, et al. "Beyond blacklists: learning to detect malicious web sites from suspicious URLs." Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009

[6] Nguyen, Luong Anh Tuan, et al. "A novel approach for phishing detection using URL-based heuristic." Computing, Management and Telecommunications (ComManTel), 2014 International Conference on. IEEE, 2014.

[7] Wikipedia. (2015. March) Uniform Resource Loactor. Avaliable: http://en.wikipedia.org/wiki/Uniform_resource_locator

[8] Kausar, Firdous, et al. "Hybrid Client Side Phishing Websites Detection Approach." International Journal of Advanced Computer Science and Applications (IJACSA) 5.7 (2014).

[9] Sunil, A. Naga Venkata, and Anjali Sardana. "A pagerank based detection technique for phishing web sites." Computers & Informatics (ISCI), 2012 IEEE Symposium on. IEEE, 2012.

[10] Mohammad, Rami M., Fadi Thabtah, and Lee McCluskey. "Intelligent rule-based phishing websites classification." Information Security, IET 8.3 (2014): 153-160.

[11] Canali, Davide, et al. "Prophiler: a fast filter for the large-scale detection of malicious web pages." Proceedings of the 20th international conference on World wide web. ACM, 2011.

[12] Xiang, Guang, et al. "Cantina+: A feature-rich machine learning framework for detecting phishing web sites." ACM Transactions on Information and System Security (TISSEC) 14.2 (2011): 21.

[13] WANG, Wei-Hong, et al. "A Static Malicious Javascript Detection Using SVM." strings. Vol. 40. 2013.

[14] L. Ladha and T. Deepa, "Feature selection methods and algorithms," International journal on computer science and engineering, vol 3, no 5, 2011.

[15] Hou, Yung-Tsung, et al. "Malicious web content detection by machine learning." Expert Systems with Applications 37.1 (2010): 55-60.

[16] Cao, Ye, Weili Han, and Yueran Le. "Anti-phishing based on automated individual white-list." Proceedings of the 4th ACM workshop on Digital identity management. ACM, 2008.

[17] Huh, Jun Ho, and Hyoungshick Kim. "Phishing detection with popular search engines: Simple and effective." Foundations and Practice of Security. Springer Berlin Heidelberg, 2012. 194-207.

[18] Abela, Kevin Joshua, et al. "An automated malware detection system for android using behavior-based analysis AMDA." International Journal of Cyber-Security and Digital Forensics (IJCSDF) 2.2 (2013): 1-11.

[19] Abu-Nimeh, Saeed, et al. "A comparison of machine learning techniques for phishing detection." Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit. ACM, 2007.

[20] OpenCV. (2015. March) Random Trees. [Online]. Available: http://docs.opencv.org/modules/ml/doc/random_trees.html

About Author (s):

Jin-Lee Lee received her Master degree in education from Dankook University of Korea and in course of PhD from Konkuk University. Her research interest lies in Security, Machine learning and anti-phishing

Dong-Hyun Kim is in course of his Bachelor of Science in computer science. His research interest lies in machine learning, anti-phishing

Chang-Hoon Lee received his PhD degree from KAIST of korea. His research interest lies in Artificial Intelligent, Security