A Survey of URL-based Phishing Detection

Eint Sandi Aung[†a)] Chaw Thet Zan[†b)] and Hayato YAMANA[†c)]

† Department of Computer Science and Communication Engineering, Graduate School of Fundamental Science and Engineering, Waseda University, Tokyo, 159-8555, Japan.

E-mail : a) eintsandiaung@toki.waseda.jp, b) chawthetzan@fuji.waseda.jp, c) yamana@waseda.jp

Abstract Cyber phishing is regarded as a theft of personal information in which phishers, also known as attackers, lure users to surrender sensitive data such as credentials, credit card and bank account information, financial details, and other behavioral data. Phishing detection is becoming a crucial research area, attracting increased focus as the number of phishing attacks grows. Furthermore, because attackers are innovating various techniques, detection has become a primary concern of developers. A number of phishing detection schemes has been built into their architecture, such as whitelist-, blacklist-, content, visual similarity and URL-based in general. Each has its individual advantages and drawbacks. In this survey paper, we emphasize on URL-based phishing detection techniques, because we consider the URL to be a significant criterium in preventing phishing attacks. Moreover, examining URL-based features can also encourage faster processing than other approaches. In this work, we aim to understand the structure of URL-based features and surveying their diverse detection techniques and mechanisms. We then analyze the performance based on the combinations of URL features on different datasets. Finally, we summarize our findings to promote better URL-based phishing detection systems.

Keyword Phishing, URL-based, Web Security, Features

1. Introduction

Phishing is a cyber threat in which attackers take advantage of users by mimicking legitimate authentic, websites in order to steal sensitive information such as passwords and bank statements. Phishing is performed through different mediums: internet, short message service and voice. Their targeted vectors can be email, instant messaging, smishing (short message phishing), vishing (voice phishing) and websites [1]. In this paper, phishing refers to web phishing through the Internet. Although phishing can be protected against by: (1) user awareness, and (2) technology-based approaches, the former cannot be completely trusted since it relies on humans–not all of whom are aware of phishing. Thus, our survey focuses on the latter for phishing detection.

According to the Anti Phishing Working Group (APWG) 3rd Quarter, 2018 report [2], the total number of detected phishing attacks was 151,014. Although this number reported has reportedly dropped since 2nd Quarter, 2018, it is still significant statistic for the public to be aware of. According to the report, there was an increase in the use of web page redirects for hiding phishing sites. As the number of page redirects used by phishers is enormously increasing, more users are lured to actual phishing sites. When users click on phishing links, they are being taken to phishing sites via other sites, where their credential information is requested. "This obfuscation technique is an effort by the phishers to hide the phishing URL – most notably from detection via web server log referrer field monitoring," said Stefanie Ellis, Anti-Fraud Product Marketing Manager at MarkMonitor [50].

Furthermore, half of the phishing sites are currently using HTTPS and SSL certificates to confuse users. PhishLabs–an APWG member that provides services against cyberattacks, reported that half of all phishing sites are using SSL encryption to deceive users with the familiar green lock symbol while some phishers even add HTTP encryption.

These occurrences illustrate that phishers have an increasing preference for URL-based attacks to gather sensitive information. Therefore, we emphasize URL-based phishing detection in our survey. Moreover, URL-based phishing detection can reduce workload and processing time compared to other approaches such as blacklist, content and visual similarity.

In this work, we survey different techniques for URL-based phishing detection. The objective of this paper is to summarize our systematic analysis of phishing detection based on URLs' characteristics–which, we believe, have a significant effect on the detection.

This paper is organized as follows: Section 2 presents a review of the literature related to phishing and its various detection categories. Section 3 consists of the architecture of URL-based phishing and a survey of the diverse features, datasets nature, methods, and evaluation metrics. Section 4 provides a summary and our opinions on existing

techniques. Section 5 presents the conclusion followed by references.

2. Literature Review

Just as phishing has various unique characteristics, so do the detection techniques and methods. However, phishing approaches can generally be classified into five categories: whitelist-, blacklist-, content-, visual similarity- and URL-based. We list an overview of each approach for a better understanding as follows:

2.1. Whitelist-Based Approach

Kang et al. [3] proposed an approach based on white-listed sites in 2007. They performed a URL similarity check to distinguish phishing sites from otherwise and a mechanism comparing with Domain Name System (DNS) query to overcome DNS pharming attacks– problem for relying on DNS from previous researches. In 2008, Cao et al. [4] also presented an automated individual white-list approach, in which the system maintains a user's previous login and warns when unfamiliar access has occurred. Although whitelist-based methods seem effective for phishing detection, there is a limitation on getting legitimate sites all on the web. An abundant list of reliable websites is necessary for a robust system with high accuracy; otherwise, false positive rates increase due to a lack of white-listed websites information, which is practically impossible to collect all legitimate sites in the world.

2.2. Blacklist-Based Approach

Web browsers–such as Google Safe Browsing – that defend against phishing attacks by updating a list of black-listed sites. In 2008, Sharifi et al. [5] proposed a new black-ist generator technique to solve the common issues of maintaining an up-to-date list. However, since their proposed system relies on third-party services (like Google) for searching domain name to compare top results, it results in poor performance. Furthermore, blacklist approaches encounter the major issue of zero-hour phishing attacks because newly created phishing sites are not in the list. PhishNet [22] also predicts phishing attacks based on a blacklist scheme. It uses five heuristics–top-level domain, IP address, directory structure, query string and brand name–for combinations of blacklists to predict new phishing sites. Although it cannot detect zero-hour phishing sites, it achieves 95% true positive rate and 3% false positive rate over large datasets.

2.3. Content-Based Approach

Zhang et al. [6] presented a novel approach, so-called CANTINA in 2007. Their work is based on Term Frequency - Inverse Document Frequency (TF-IDF) information retrieval algorithm used to detect phishing websites. CANTINA alone resulted in a high false positive rate due to limitations on the number of search engine results. This means that as they increase the number of results, false positive rate will decrease while true positive rate remains the same, which is not optimal. Thus, they used several heuristics to reduce the false positive rate and improve accuracy. Their approach achieved a better outcome compared to popular anti-phishing toolbars, achieving 97% true positive and 1% false positive rate. In 2011, Xiang et al. [7] further improved CANTINA, calling it CANTINA^{+}, which is regarded as the most comprehensive feature-rich approach in content-based phishing detection. It achieved a better 0.4% false positive rate and over 92% true positive rate. However, since both approaches use search engines and third-party services, DNS compromising became a challenging threat. Similar works can be found in [23][24][25].

2.4. Visual-Similarity-Based Approach

Wenyin et al. [8] proposed a simple visual-similarity-based approach in 2005. Their system performed phishing detection on three levels of similarity matrices; (i) block-level similarity, (ii) layout-similarity and (iii) overall-style similarity. However, the most representative work on visual similarity was later presented by Fu et al. [9] in 2006 using the Earth Mover Distance (EMD). EMD was used to calculate the signatures of two images for visual similarity. Although their method performed well in accuracy with 89% true positive and 0.71% false positive rates, the significant workload required to process two images was a performance drawback, compared to other approaches.

Chen et al. [16] introduced a heuristic anti-phishing system to model perceptual similarity. They employed a logistic regression algorithm for normalizing page content features. Although the proposed method achieved 100% true positive rate, it had 0.74% false positive rate, which could be improved. There are many similar works based on visual similarity including [27] [28] [29] [30] [31] [32] [33] [34] [35] [36] [37].

2.5. URL-Based Approach

M. Aburrous, M. A. Hossain, K. Dahal and F. Thabtah [10] proposed and intelligent phishing detection system for e-banking using fuzzy data mining in 2010. The experiment was performed based on fuzzy logic with data mining algorithms. They showed how effective URL-based approaches are for phishing detection. Overall, URL-based methods perform faster than any other, including content- and visual-similarity based approaches. More importantly,

they work well on zero-hour phishing attacks, which are becoming a major concern in modern anti-phishing society. In upcoming sections, we further discuss details of URL-based detection. Similar works can be found in [18] [19].

2.6. Other Approaches

A variety of alternative techniques are used by researchers in phishing detection. Such techniques include heuristic[17], hybrid[13], machine-learning[20], DNS-based and others [21][26]. Additionally, several surveys regarding different schemes are performed by researchers [11][12][15][16].

3. Architecture of URL-Based Phishing

URL-based phishing attacks are mainly performed by embedding sensitive words or characters in a link that:

1. Mimic similar but misspelling words.
2. Contain special characters for redirecting.
3. Use shortened URLs.
4. Use sensitive keywords which seem reliable.
5. Add a malicious file in the link and so on.

Figure 1 shows how URL phishing is performed. When phishers mimic as reliable sites, users submit credential information to attackers without knowing the website is faked.

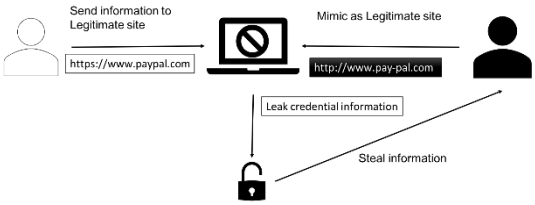

Fig 1. Architecture of URL phishing

3.1. URL Types

According to PhishStorm [11], five different types of URL obfuscation are employed. URLs are obfuscated by mixing keywords in paths, queries and low-level domains as follows listed with examples in Table 1:

Type 1: Obfuscation with other domains
Type 2: Obfuscation with keywords
Type 3: Typo-squatting domains
Type 4: Obfuscation with IP address
Type 5: Obfuscation with URL shorteners

Table 1. URL Obfuscation Types

Type	Sample
Type 1	http://school497.ru/222/www.paypal.com/29370274276105805
Type 2	http://quadrodeofertas.com.br/www1.paypal-com/encrypted/ssl218
Type 3	http://cgi-3.paypal-secure.de/info2/verikerdit.html
Type 4	http://69.72.130.98/javaseva/https://paypal.com/uk/onepagepaypal.htm
Type 5	http://goo.gl/HQx5g

3.2. URL-Based Detection Schemes

Previous research has primarily focused on two detection schemes as:

- Algorithms-based (Feature Extraction + Classification)
- Feature Engineering-based (Feature Extraction + Feature Selection + Classification)

We discuss these schemes in the following sections.

3.2.1. Commonly Used Algorithms

Here, we list the machine learning algorithms most commonly used in phishing detection literature.

3.2.1.1. Naïve Bayes

Naïve Bayes (NB) is a simple, yet effective classifier used in numerous applications. In a NB classifier, x is the features vectors, $y \in \{0,1\}$ is a label representing either a phishing or legitimate website ($y = 1$ for phishing and $y = 0$ for legitimate), and $P(x|y)$ is the conditional probability of the feature vector given its label. Assuming phishing and legitimate websites are equally probable, the posterior probability of x belongs to y=1 is as follows:

$$P(y = 1|x) = \frac{P(x|y=1)}{P(x|y=1)+P(x|y=0)} \qquad (1)$$

3.2.1.2. Support Vector Machine

Support Vector Machine (SVM) is a typical machine-learning method for classification and regression. SVM finds the optimal separating hyperplane between two labels. It can be expressed by the kernel function $K(x,x')$, in which the similarity of two feature vectors is computed, and non-negative coefficients $\propto_i$. SVM indicates which training examples lie closely to the decision boundary. It classifies data by computing distance to decision boundary.

$$h(x) = \sum_1^n \propto_i (2y_i - 1)\, K(x_i, x) \qquad (2)$$

3.2.1.3. Random Forest

A Random Forests is built with random attribute selection using bagging. Random Forests employ a divide and conquer approach (ensemble mechanism) for improving performance. In a random forest, the mechanism combines various random subsets of trees. The overall result is calculated based on the average, or weighted average, of the individual results. The accuracy depends on a measure of the dependence between the classifier and the strength of the individual classifiers and they improve the problem of overfitting of the decision trees.

3.2.1.4. Convolutional Neural Network

Convolutional Neural Network (CNN) is a category of deep neural networks used to analyze image processing. CNN requires relatively little pre-processing compared to other image classification algorithms. It learns features

themselves – a major advantage in feature engineering, which is different from other classifications pre-specified by researchers in traditional phishing detection. As it is mostly designed for image classification, it is performed on character-level embeddings for phishing detection. CNN networks contain a convolution layer, pooling layer and fully connected network with non-linear activation function. Table 2 lists several algorithms commonly used in the phishing detection field.

Table 2. Commonly Used Algorithms

No	Algorithms	References
1	Naïve Bayes	[39][41][43]
2	Logistic Regression	[39][43][48]
3	Random Forest	[14][39][40][43][47][48]
4	Support Vector Machine	[39][41][43][44]
5	k-means	[44]
6	Neural Network	[38][39][45][46]
7	LSTM	[40][44][45][46]
8	Decision Tree	[47][48]

3.2.2. Common URL-Based Features

In the feature engineering field of phishing detection, researchers apply several features depending on their detection techniques. We survey the more commonly used features in URL-based detection in Table 3.

Table 3. Common URL-Based Features

No	Feature Name	Description
1	IP address	Check if IP address is presented in existing domains
2	Avg. words length	Count average length of meaningful words in entire domain name
3	exe/zip	Check if exe/zip is present in URL
4	No of dots	Count # of dots in URL
5	Special symbols	Count special symbols in URL
6	URL length	Count # of characters in URL
7	Top-level domain (TLD) feature	Validate TLD-based features [39][40][44]
8	“http” count	Count # of “http” in URL
9	Brand name	Extract brand name in URL domain
10	“//” redirection	Check if “//” is included in URL path
11	Domain separated by “-“	Check if “-“ is included in domain name
12	Multi-sub domain	Check how many # of multi-subdomains are included in URL
13	Suspicious words	Check if suspicious words are included in URL
14	Digits in domain	# of digits in domain
15	Character entropy	Calculate character distribution in entire URL using entropy
16	Shorten URL	Check if URL is shortened

3.3. Evaluation Matrices

Here, we assume that N represents the number of legitimate/phishing websites and P represents phishing and L represents legitimate.

True Positive rate (TPR): the ratio of the number of correctly classified phishing attacks ($N_{P\to P}$) to the total number of phishing attacks ($N_{P\to P}+N_{P\to L}$). See Equation (3) for details.

False Positive rate (FPR): the ratio of the number of legitimate sites that are incorrectly detected as phishing attacks ($N_{L\to P}$) to the total number of all existing legitimate sites ($N_{L\to L}+N_{L\to P}$). See Equation (4) for details.

True Negative rate (TNR): the ratio of the number of correctly classified legitimate sites ($N_{L\to L}$) to the total number of existing legitimate sites ($N_{L\to L}+N_{L\to P}$). See Equation (5) for details.

False Negative rate (FNR): the ratio of the number of phishing attacks that are incorrectly classified as legitimate ($N_{P\to L}$) to the total number of phishing attacks ($N_{P\to P}+N_{P\to L}$). See Equation (6) for details.

Precision (P): the ratio of correctly detected phishing attacks ($N_{P\to P}$) to the total number of attacks detected as phishing ($N_{L\to P}+N_{P\to P}$). See Equation (7).

Recall (R): equivalent to TP rate. See Equation (8).

Accuracy (ACC): the ratio of the sum of correctly classified phishing and legitimate sites ($N_{L\to L}+N_{P\to P}$) to the total sites ($N_{L\to L}+N_{L\to P}+N_{P\to P}+N_{P\to L}$). See Equation (9) in details.

$$TPR=\frac{N_{P\to P}}{N_{P\to P}+N_{P\to L}} \quad (3)$$

$$FPR=\frac{N_{L\to P}}{N_{L\to L}+N_{L\to P}} \quad (4)$$

$$TNR=\frac{N_{L\to L}}{N_{L\to L}+N_{L\to P}} \quad (5)$$

$$FNR=\frac{N_{P\to L}}{N_{P\to P}+N_{P\to L}} \quad (6)$$

$$P=\frac{N_{P\to P}}{N_{L\to P}+N_{P\to P}} \quad (7)$$

$$R=TP \quad (8)$$

$$ACC=\frac{N_{L\to L}+N_{P\to P}}{N_{L\to L}+N_{L\to P}+N_{P\to P}+N_{P\to L}} \quad (9)$$

We surveyed and listed URL-based phishing detection mechanisms. Table 4 describes the comparative evaluation results based on the above matrices.

3.4. Datasets Nature

Researchers collect data sources from popular websites

such as Alexa and DMOZ for legitimate, and PhishTank and OpenPhish for phishing. Several common sources are listed in Table 5.

Table 5.Data Sources

Type	Data Source
Legitimate	digg58.com, Alexa, DMOZ, payment gateway, Top banking website
Phishing	PhishTank, OpenPhish, VirusTotal, MalewareDomainList, MalewareDomains, jwSpamSpy

We discover that a majority of the researches focuses on imbalanced data as the number of phishing sites cannot be compared with that of legitimate sites. However, several studies use balanced datasets to avoid dataset bias.

4. Summary and Opinion

In our survey perspective, we observed two perspectives from the existing detection schemes; (i) dataset perspective, and (ii) feature perspective.

From the dataset perspective, researchers primarily analyze the detection method on imbalanced data, in which the majority class is legitimate sites. This results in a biasing majority class. Put differently, the result is biased although it has a high false positive rate. To address this, oversampling on minority data becomes effective since it balances data size by realistic automated minority-class data.

From the feature perspective, we find that several URL-based features–such as the number of subdomains and URL length could also be biased since they highly rely on the dataset. In other words, many researchers use Alexa.com for legitimate dataset, in which only index pages of highly ranked websites are provided. However, phishing datasets from PhishTank.com or OpenPhish.com list the entire URLs of the phishing webpages in which phishers use free hosting services that are highly ranked in Alexa. Thus, as for the number of subdomains, legitimate sites from Alexa.com will not have any, while phishing sites will. Furthermore, phishers have complete control over URL composition except for the domain name. Features like URL length can be easily manipulated. Therefore, researchers have recently targeted domain name-based features–instead of entire URL–to extract characteristics of domain name and current page content.

5. Conclusion

In this paper, we described our systematic survey of existing URL-based phishing detection techniques from different views. Although previous survey papers exist, they generally focus on overall phishing detection techniques, while we focused on detailed URL-based detection with respect to features. Firstly, we reviewed the literature on overall phishing detection schemes. Second, we discussed the architecture of URL-based phishing, and commonly used algorithms and features. Third, common data sources were listed, and comparative evaluation results and matrices were shown for better survey. Finally, we concluded with our recommendations for more effective phishing detection in the future.

References

[1] K. L. Chiew, K. S. C. Yong, C. L. Tan, “A survey of phishing attacks: their types, vectors and technical approaches,” in Expert Systems with Applications, vol.106, pp.1-20, 2018.

[2] I. Tanaka and J. Suzuki, “Web and Database Technologies”, Proc. of ACM SIGMOD, pp. 10-22, 201 APWG, “Phishing activity trends report, 3rd Quarter2018,”Internet:http://docs.apwg.org/reports/apwg_trends_report_q3_2018.pdf, Dec. 12, 2018 [Dec. 25, 2018]

[3] J. Kang and D. Lee, “Advanced white list approach for preventing access to phishing sites,” Proc. International Conference on Convergence Information Technology (ICCIT 2007), pp.491-496, 2007.

[4] Y. Cao, W. Han, and Y. Le, “Anti-phishing based on automated individual white-list,” Proc. the 4th ACM Workshop on Digital Identity Management, pp.51–60, 2008.

[5] M. Sharifi and S. H. Siadati, “A phishing sites blacklist generator,” in IEEE/ACS International Conference on Computer Systems and Applications, pp. 840-843, 2008.

[6] Y. Zhang, J. I. Hong, L. F. Cranor, “Cantina: a content-based approach to detecting phishing web sites,” Proc. the 16th International Conference on World Wide Web, pp.639-648, 2007.

[7] G. Xiang, J. Hong, C. P. Rose, L. Cranor, “Cantina+: a feature-rich machine learning framework or detecting phishing web sites,” in ACM Transactions on Information and System Security (TISSEC), vol.14, no.2, pp.21:1-21:28, 2011.

[8] L. Wenyin, G. Huang, L. Xiao Yue, Z. Min, X. Deng, “Detection of phishing webpages based on visual similarity,” in Special interest tracks and posters of the 14th International Conference on World Wide Web, pp. 1060-1061, 2005.

[9] Y. Fu, L. Wenyin and X. Deng, "Detecting phishing web pages with visual similarity assessment based on earth mover's distance (EMD)," in IEEE Transactions on Dependable and Secure Computing, vol. 3, no. 4, pp. 301-311, 2006.

[10] M. Aburrous, M. A. Hossain, K. Dahal, F. Thabtah, “Intelligent phishing detection system for e-banking using fuzzy data mining,” in Journal of Expert System with Applications, vol.37, no.12, pp.7913-7921, 2010.

[11] S. Marchal, J. Francois, R. State, T. Engel, “PhishStorm: detecting phishing with streaming analytics,” in IEEE Transactions on Network and Service Management, vol.11, no.4, pp.458-471, 2014.

[12] M. Khonji, Y. Iraqi, A. Jones, “Phishing detection: a literature survey,” in IEEE Communications Surveys & Tutorials, vol.15, no.4, pp.2091-2021, 2013.

[13] M. Dadkhah, S. Shamshirband, A. Wahab, “A hybrid approach for phishing web site detection,” in the

Table 4. Comparison of Evaluation Results in Existing Phishing Detection

Paper	Performance							Data Set					URL Features	Algorithms	Year
	Acc	TP	FP	Prec	Rec	F1 Score	AUC	Dataset source		Dataset size		Dataset type			
								Legitimate	Phishing	Legitimate	Phishing				
[38]							99.29	VirusTotal	VirusTotal	16M	0.9M	Imbalanced	Primary domain feature, path feature, file extension feature (*3)	CNN	2017
[39]	99.09							Alexa+ Payment Gateway+ Top Banking Website	PhishTank+ OpenPhish	1600+ 66+ 252	1528+ 613		No. of dots, Special symbols, URL length, special words, position of TLD, HTTP count, brand name, data URI (*8)	RF SVM NN LR NB	2017
[40]	98.76			98.60	98.93	98.76	99.91	Common Crawl	PhishTank	1M	1M	Balanced	Path length, URL entropy, length ratio, '@' and '-' count, punctuation count, TLDs count, IP address, suspicious words count, Euclidean distance, Kolmogorov-Smirnov statistic (*10)	RF LSTM	2017
[41]	95.80							Alexa+ Search Engine	PhishTank	500+ 500	1000	Balanced	URL size, no. of hyphens, no. of dots, no. of numeric characters, IP address, similarity index (Levenshtein, Jaro Winkler, Normalized Levenshtein, longest common subsequence, Q Gram, Hamming) (*6)	NB Bayes SVM	2018
[42]	95.00								PhishTank				IP Address, redirection of page using "//", adding prefix or suffix separated by "-", subdomain and multi-subdomain, URLs having @ symbol (*6)	IG Ranker Method	2018
[43]		99.70	0.40	99.70	99.70	99.70	1.00	digg58.com+ GitHub	PhishTank+ GitHub	16516+ 37,667	12483+ 24,905	Slightly imbalanced	IP address, no. of dots, no. of "/", special characters, abnormal length of domain, character distribution (*36)	RF MLP NB LR J48 SVM	2018
[44]							70.10- eBay 71.01- PayPal 70.10- BoA 97.65- Sorio	eBay+ PayPal+ Bank of America + Sorio et al.	eBay+ PayPal+ Bank of America+ Sorio et al.	18800+ 17572+ 9408+ 82101	8529+ 9690+ 4610+ 6562	Imbalanced	No. usage of domain name, URL length, domain separated by "-", multiple subdomains, usage of "@" symbol, no. of TLD in the path, no. of suspicious words, no. of punctuation symbols used, digits in domain, entropy, Kullback-Leibler divergence, no. of "-" in path, vowel/consonant ratio, digit/letter ratio, usage of brand names, long	SVMSMOTE bSMOTE2 RMR ADASYN	2018

													hostnames, short hostnames, no. of ":" in hostname (*18)		
[45]	96.89							Search Engine+ Common Crawl+ Twitter Stream API	PhishTank	456300			Features from [49][38]	ANN LSTM	2018
[46]	D1: 99.47 D2: 99.92								PhishTank+ OpenPhish+ Malware Domain List+ Malware Domains	Data set1: 90101 Data set2: 26000				CNN CNN-LSTM Bigram	2018
[47]	99.44							Alexa	PhishTank + Malware Domain List+ Spam Domain List jwSpamSpy	26041	26041	Balanced	117 static and dynamic features	J48 Simple Cart RF RT ADTree REPTree Majority Voting	2018
[48]	97.70	98.30	2.60					Alexa+ Network Security Challenge	PhishTank	3305	2892		IP address, suspicious characters, network protocol, alexa ranking, length of entire URL, length of hostname, length of main domain name, no. of dots in hostname, no. of dots in URL path, URL token count, hostname token count, search engine result (*12)	KNN LR RF DT GBDT XGBST DF	2018

(*) represents number of URL-based feature

Electronic Library, vol.34, no.6, pp.927-944, 2016.

[14] A. Subasi, E. Molah, F. Almkallawi, T. J. Chaudhery, "Intelligent phishing website detection using random forest classifier," in the International Conference on Electrical and Computing Technologies and Applications (ICECTA), pp.1-5, 2017.

[15] Z. Duo, I. Khalil, A. Khreishah, A. Al-Fuqaha, M. Guizani, "Systematization of knowledge (SoK): a systematic review of software based web phishing detection," in IEEE Communications Surveys & Tutorials, vol.19, no.4, pp.2797-2819, 2017.

[16] A. Oest, Y. Safei, A. Doupe, G. J. Ahn, B. Wardman, G. Warner, "Inside a phisher's mind: understanding the anti-phishing eco system through phishing kit analysis," Proc. 2018 APWG Symposium on Electronic Crime Research (eCrime), 2018.

[17] M. Babagoli, M. P. Aghababa, V. Solouk, "Heuristic nonliner regression strategy for detecting phishing websites," in Soft Computing, vol.22, no.236, pp.1-13, 2018.

[18] E. Sorio, A. Bartoli, E. Medvet, "Detection of hidden fraudulent URLs within trusted sites using lexical features," in 2013 International Conference on Availability, Reliability and Security, 2013.

[19] M. N. Feroz, S. Mengel, "Phishing URL detection using URL ranking," Proc. 2015 IEEE International Congress on Big Data, pp.635-638, 2015.

[20] A. Hodzic, J. Kevric, "Comparison of machine learning techniques in phishing website classification," Proc. International Conference on Economic and Social Studies (ICESoS'16), vol.3, pp.249-256, 2016.

[21] M. Moghimi, A. Y. Varjani, "New rule-based phishing detection method," in Expert System with Applications, vol.53, pp.231-242, 2016.

[22] P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, "Phishnet: predictive blacklisting to detect phishing attacks," Proc. IEEE INFOCOM, 2010, pp.1-5, 2010.

[23] G. Xiang and J. I. Hong, "A hybrid phish detection approach by identity discovery and keywords retrieval," Proc. the 18th International Conference on World Wide Web, pp.571-580, 2009.

[24] V. Ramanathan and H. Wechsler, "Phishing website detection using latent dirichlet allocation and adaboost," in 2012 IEEE International Conference on Intelligence and Security Informatics (ISI), pp.102-107, 2012.

[25] M. Aburrous and A. Khelifi, "Phishing detection plug-in tool-bar using intelligent fuzzy-classification mining techniques," in the International Journal of Soft Computing and Software Engineering [JSCSE], vol.3, no.3, pp.54-61, 2013.

[26] T. Chen, T. Stepan, S. Dick, and J. Miller, "An anti-phishing system employing diffused information," in Journal of ACM Transactions on Information and System Security (TISSEC), vol.16, no.4, pp.1-31, 2014.

[27] Y. Fu, W. Liu, and X. Deng, "EMD based visual similarity for detection of phishing webpages," Proc. of International Workshop on Web Document Analysis, 2005.

[28] L. Wenyin, G. Huang, L. Xiaoyue, X. Deng, and Z. Min, "Phishing web page detection," in the 8th International Conference on Document Analysis and Recognition (ICDAR'05), pp.560-564, 2005.

[29] Masanori Hara, Akira Yamada, and Yutaka Miyake, "Visual similarity-based phishing detection without victim site information," in IEEE Symposium on Computational Intelligence in Cyber Security (CICS'09), pp.30-36, 2009.

[30] G. Liu, B. Qiu, and L. Wenyin, "Automatic detection of phishing target from phishing webpage," in IEEE 20th International Conference on Pattern Recognition (ICPR), pp.4153-4156, 2010.

[31] S. Afroz and R. Greenstadt, "Phishzoo: an automated web phishing detection approach based on profiling and fuzzy matching," Proc. of the 5th IEEE International Conference on Semantic Computing (ICSC), 2009.

[32] C. Kuan-Ta, C. Jau-Yuan, H. Chun-Rong, and C. Chu-Song, "Fighting phishing with discriminative keypoint features," in IEEE Internet Computing, vol.13, no.3, pp.56-63, 2009 .

[33] H. Masanori, Y. Akira, and M. Yutaka, "Visual similarity-based phishing detection without victim site information," in IEEE Symposium of Computational Intelligence in Cyber Security, CICS'09, pp.30-36, 2009.

[34] C. Teh-Chung, D. Scott, and M. James, "Detecting visually similar web pages: application to phishing detection," in ACM Transactions on Internet Technology (TOIT), vol.10, no.2, pp.1-5, 2010.

[35] Nuttapong Sanglerdsinlapachai and Arnon Rungsawang, "Using domain top-page similarity feature in machine learning-based web phishing detection," in the 3rd International Conference on Knowledge Discovery and Data Mining, WKDD'10, pp.187-190, 2010.

[36] H. Zhang, G. Liu, T. Chow, and W. Liu, "Textual and visual content-based anti-phishing: a bayesian approach," in IEEE Transactions on Neural Networks, vol.22, no.10, pp.1532-1546, 2011.

[37] Max-Emanuel Maurer and Dennis Herzner, "Using visual website similarity for phishing detection and reporting," in ACM Conference on Extended Abstracts on Human Factors in Computing Systems, pp.1625-1630, 2012.

[38] H. Le, Q. Pham, D. Sahoo, S. C. H. Hoi, "URLNet: learning a url representation with deep learning for malicious url detection," in Journal of Computing Research Repository (CoRR), 2018.

[39] A. K. Jain and B. B. Gupta, "Towards detection of phishing websites on client-side using machine learning based approach," in Journal of Telecommunication System, vol.64, no.4, pp.687-700, 2017.

[40] A. C. Bahnsen, E. C. Bohorquez, S. Villegas, J. Vargas and F. A. González, "Classifying phishing URLs using recurrent neural networks," Proc. 2017 APWG Symposium on Electronic Crime Research (eCrime), pp.1-8, 2017.

[41] M. Zouina and B. Outtaj, "A novel lightweight URL phishing detection system using SVM and similarity index," in Journal of Human-centric Computing and Information Sciences, vol.7, no.17, pp.1-13, 2017.

[42] S. Parekh, D. Parikh, S. Kotak and P. S. Sankhe, "A New Method for Detection of Phishing Websites: URL Detection," in 2018 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT), pp.949-952, 2018.

[43] C. Liu, L. Wang, B. Lang and Y. Zhou, "Finding Effective Classifier for Malicious URL Detection," Proc. the 2nd International Conference on Management Engineering, Software Engineering and Service Sciences (ICMSS), pp.240-244, 2018.

[44] A. Ankesh, G. Kshitij, M. Joel, P. Noseong, C. Tanmoy, C. Bei-Tseng, "Phishing URL detection with oversampling based on text generative adversarial networks," Proc of 2018 IEEE International Conference on Big Data (Big Data), pp.1167-1176, 2018.

[45] S. Shivangi, P. Debnath, K. Saieevan and D. Annapurna, "Chrome extension for malicious URLs detection in social media applications using artificial neural networks and long short term memory networks," in the International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1993-1997, 2018.

[46] A. Vazhayil, R. Vinayakumar and K. Soman, "Comparative study of the detection of malicious URLs using shallow and deep networks," in the 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2018, pp.1-6, 2018.

[47] R. P. Dharmaraj, J. B. Patil, "Malicious URLs detection using decision tree classifiers and majority voting technique," in Journal of Cybernetics and Information Technologies, vol.18, no.1, pp.11-29, 2018.

[48] H. Yuan, X. Chen, Y. Li, Z. Yang and W. Liu, "Detecting phishing websites and targets based on URLs and webpage links," in the 24th International Conference on Pattern Recognition (ICPR), pp. 3669-3674, 2018.

[49] D. Sahoo, C. Liu, and S. Hoi, "Malicious url detection using machine learning: A survey," in Computing Research Repository (CoRR), 2017.

[50] APWG Report of Criminal Innovation Ramps Up with Phishing Attack in 2018 by Business Wire, Internet:https://www.businesswire.com/news/home/20181212005652/en/APWG-Report-Criminal-Innovation-Ramps-Phishing-Attacks, Dec. 12, 2018 [Jan. 8, 2019]