International Journal of Engineering & Technology, 7 (4.31) (2018) 7-14

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Building Standard Offline Anti-phishing Dataset for Benchmarking

Kang Leng Chiew[1]*, Ee Hung Chang[2], Choon Lin Tan[3], Johari Abdullah[4], Kelvin Sheng[5] Chek Yong[6]

Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
**Corresponding author E-mail: klchiew@unimas.my*

Abstract

Anti-phishing research is one of the active research fields in information security. Due to the lack of a publicly accessible standard test dataset, most of the researchers are using their own dataset for the experiment. This makes the benchmarking across different anti-phishing techniques become challenging and inefficient. In this paper, we propose and construct a large-scale standard offline dataset that is downloadable, universal and comprehensive. In designing the dataset creation approach, major anti-phishing techniques from the literature have been thoroughly considered to identify their unique requirements. The findings of this requirement study have concluded several influencing factors that will enhance the dataset quality, which includes: the type of raw elements, source of the sample, sample size, website category, category distribution, language of the website and the support for feature extraction. These influencing factors are the core to the proposed dataset construction approach, which produced a collection of 30,000 samples of phishing and legitimate webpages with a distribution of 50 percent of each type. Thus, this dataset is useful and compatible for a wide range of anti-phishing researches in conducting the benchmarking as well as beneficial for a research to conduct a rapid proof of concept experiment. With the rapid development of anti-phishing research to counter the fast evolution of phishing attacks, the need of such dataset cannot be overemphasised. The complete dataset is available for download at http://www.fcsit.unimas.my/research/legit-phish-set.

Keywords: *Anti-phishing; Dataset for benchmarking; Features; Legitimate and phishing webpages*

1. Introduction

The advancement of information technology has provided many benefits to our life as we are able to handle many daily works by using the Internet services. For example, instead of going to the respective service counter, people nowadays are able to pay their bills at any place they feel convenient and with an Internet connection. However, the extension of this convenience has also come along with some immoral activities that are known as the online crimes. Online criminals always gained their illegal profits from their targets through the vulnerability of the Internet service, and one of the common online crimes is called online phishing.

Online phishing is a security threat which combines social engineering and website spoofing techniques to deceive users into revealing their confidential information [1, 2]. Typically, phisher will try to harvest online users credential such as username, passwords and credit card detail by masquerading as a trustworthy entity on the Internet [3, 4, 5]. To prevent the users from becoming the victims of phishing attacks, many software vendors, research institute and companies have released various anti-phishing techniques [6].

There are many survey publications related to the phishing attacks and anti-phishing techniques. However, according to the best of our knowledge, survey publication on the anti-phishing dataset is still unavailable. The discussions on the correlation of anti-phishing dataset and the experimental results are still inadequate and not profound. Furthermore, there is a lack of a widely recognised standard offline dataset which available for the research community to utilise. A complete package of downloadable anti-phishing datasets is also limited. This situation has caused difficulty to create a consistent condition for fair and rapid benchmarking.

This paper aims to fulfil the gap by providing a standard offline dataset for the anti-phishing research community. Although in a different field, similar works on constructing a standard dataset can be found in [7, 8, 9]. We will look into multiple types of anti-phishing approaches from the past, review on their approaches and datasets, and identify the related anti-phishing features. This insight will ensure the dataset built later to be at optimum flexibility and adaptable by the research community. The discussion on “how a good anti-phishing offline dataset should be designed” will be included in this paper by highlighting some of the factors that may influence the accuracy of experimental results.

The remainder of the paper is structured as follows. In the next section, we will review some of the past anti-phishing works in Section 2 and discuss the dataset used in those works in Section 3. We will later discuss on the factors that may contribute to a good design of an anti-phishing offline dataset in Section 4. In Section 5, we will discuss the construction of the offline dataset in detail based on the review from previous sections. The paper concludes in Section 6.

2. Anti-Phishing Approaches and Features

There are varieties of anti-phishing techniques available in the literature and broadly they can be divided into list-based and heuristic-based approaches. Each has its own effective features to be utilised for the phishing detection. In this section, we will group and highlight the major features from the past research works according to their approaches.

2.1. List-Based

According to Zhang et al. [10], blacklisting appears to be one of the popular techniques in the anti-phishing community. Many popular web browsers have integrated this technique to detect the phishing websites [11, 12]. In this technique, a query website is checked with a list (i.e., a list of known phishing URLs), which is compiled and maintained by some consortium or organisation.

A method proposed by Cao et al. is one of the whitelisting examples that will maintain and store a whitelist at the client side automatically [13]. Prakash et al. later introduce a more dynamic and flexible list-based approach, called PhishNet [14]. This method will generate multiple variations URLs based on the existing blacklist, and the generated URLs will be served as a predictive blacklist. Their dataset can be concluded as URL oriented, as this type of method only utilises the webpage URLs.

2.2. Heuristic-Based

Another prominent approach that draws a great attention is the heuristic-based approach. This approach can overcome the limitation of list-based approach by avoiding the high effort in maintaining the up-to-date lists. The heuristic-based approach analyses a query website and extract some discriminative properties as the features for further processing to determine the legitimacy. Since heuristic-based approach covers a huge range of anti-phishing method, it can be further categorised into content-based, URL-based, and visual-based methods.

A popular example of content-based method is called CANTINA [15]. This method calculates the Term Frequency-Inverse Document Frequency (TF-IDF) from the content of a website and generates a lexical signature, which will later be used as the keywords list for search engine query. Based on the returned result, CANTINA will determine the legitimacy of the query website. Xiang and Hong [16] later enhanced the CANTINA keywords-retrieval methodology by implementing the identity-based detection algorithm. This method will first utilise two textual objects from the Document Object Model (DOM) (i.e., the title and copyright fields) and employ a technique called Named Entity Recognition (NER) to determine the identity. The main features from these methods (i.e., TF-IDF, DOM and hyperlinks) are retrieved via HTML analysis; hence the HTML file will be the major component for their dataset. Beside using the DOM feature to retrieve textual objects, it can be used in a DOM comparison technique as done by Cui et al. [17].

Another content-based method has been proposed by Liu et al. [18], where the method can determine the identity of the targeted legitimate website when a phishing webpage is detected. This method is based on the idea of a self-organised semantic data model, called Semantic Link Network (SLN). This method extracts a series of textual element like hyperlinks, keywords and textual contents, and then processes with different detection techniques (e.g., link relation, search relation, text relation and SLN) to obtain the results. APG (Anti-Phishing Gateway) is a gateway-side solution, which focuses on the path between the user's browser and the server of query webpage. APG will analyse every URL through the gateway and selectively fetch the webpage packets, and evaluate them. Zhang et al. [19] introduce another content-based phishing detection system that is based on APG called BUPT-APG. This method will deploy an adaptive cosine similarity to calculate the similarity between the generated template in the repository and the query webpage. The similarity results will determine the legitimacy of the query webpage. In addition, [20, 21, 22] are the works that contribute to the content-based method, and the major component for their dataset would be the HTML file and the webpage URL. Jain and Gupta [23] proposed a content-based method that uses the hyperlinks from the source code of a webpage as the feature for their phishing detection algorithm. This is also done by Rao and Pais [24] with the addition of checking for the presence of login form and iframe.

URL-based method is also one of the most common anti-phishing methods in heuristic approach. For example, Ma et al. [25] publish a paper describing their research on identifying phishing by examining the characteristics of the URLs and website hosting information. Garera et al. [26] introduce a method based on URL analysis to determine various patterns that are always exploited by phishers. The method utilises a logistic regression model with the extracted features to detect phishing URLs. The extracted features include page rank, domain name, URL type, and suggestive word tokens (e.g. signin, login). Hu et al. [27] used the popularity and performance of the web domains such as citation ranking, backlinks, and page rank of the web domains in their phishing detection model.

The method proposed in [28] appears to be another interesting URL-based method, which utilises spelling recommendations from the search engine and the string similarity algorithm. Recently, Sanance et al. proposed to utilise machine learning and web mining-based approach in an URL-based analysis phishing detection [29]. The method will extract a number of features from the URLs (e.g., textual properties, WHOIS information, page ranking). After that it will apply random forest and content-based algorithms to distinguish the phishing from the legitimate websites. Similar methods that belong to the URL-based method can be found in [30, 31, 32, 33, 34, 35]. The major component of the dataset for this method is webpage URL, and some third-party information such as WHOIS, page rank as the supported items. Yan et al. [36], Marchal et al. [37], Jabri and Ibrahim [38], Gupta and Shukla [39], and Singh et al. [40] used both content-based and URL-based in their phishing detection model.

Another interesting method belonged to heuristic approach is visual-based method, where it uses the visual similarity measurement to detect phishing webpages. For example, Liu et al. [41] proposed a series of visual approaches in phishing classification. This method analyses the HTML webpages and decomposes them into salient blocks, and then calculates the similarities indicated by three metrics: block-level (detail), layout (global), and style (overall). Fu et al. [42] also proposed a visual-based method that utilises Earth Movers Distance (EMD) to calculate the webpage visual similarity for phishing detection. Rao and Ali [43] used the same visual-based method as well.

Another method related to the visual-based approach is GoldPhish [44]. GoldPhish will extract all the textual contents by performing optical character recognition (OCR) on the webpage screenshot and the extracted text will be fed into the Google search engine. The legitimacy of a query webpage is depended on the matching of the search results. Recently, Choo et al. [45] proposed an approach which utilises the website favicon to evaluate the legitimacy of a website. The favicon will be used as a query image to feed into Google Image Search engine. The Google Image Search engine is a content-based image retrieval (CBIR) system that will return a list of visually similar images and related information regarding the query image. The authors at the end will perform the latent semantic analysis based on the search results to determine the legitimacy of query webpage. The visual-based methods require more raw-elements (e.g., screenshot, favicon, image and complete rendered HTML page) in their dataset for feature extraction. Researches in [46, 47, 48] are all belonged to the visual-based method.

3. Dataset Used in Existing Anti-Phishing Works

Table 4 summarised all the datasets used in various anti-phishing works mentioned in Section 2. The summary includes (i) Dataset Size, (ii) Source, and (iii) Downloadable. The Dataset Size means the total number of legitimate and phishing webpages that are used as dataset, while the Source column shows where the authors obtained their legitimate and phishing webpages. The Downloadable column indicates the availability to the public access. The "-" sign indicates no information is available to the respective publication.

From Table 4, we obtained the following observations:

- The range of the datasets is scattered from less than 100 to more than a million. The sample size differences between the legitimate and phishing in some datasets are huge. We can conclude that there is still no strong agreement on the standard size of a dataset used in the anti-phishing community.
- PhishTank is the most popular source for the research community to obtain their phishing datasets. More than 80% of the publications listed in Table 4 utilise PhishTank as their main phishing source. For the legitimate datasets, all of them are obtained from the popular web directories (e.g., Google [49], Yahoo [50], Alexa [51] and DMOZ [52]).
- Some researches are only using datasets that are at top ranking, or popular websites for the experiments (e.g., [13], [18], [44], [45], and [48]). This shows the less popular websites have been marginalised and overlooked. Lack of unpopular legitimate website will definitely cause false positive detection.
- Most of the listed publications do not have a complete description on their dataset characteristics, such as the variety of webpage languages, detail information on screenshot images and the distribution and the categorisation of webpage (i.e., the percentage for each field of website such as social media, banking and others within the dataset). Some datasets contain a high volume of repeated same popular brand which may cause bias to the experimental result. Such datasets will lower the credibility of the experimental results and may draw advantages to certain methods.

4. Influencing Factor on Dataset Design

The dataset is a collection of various elements of webpages downloaded from different websites. In order to build a standard offline dataset that is suitable for a wide range of experiments, we propose to consider the following factors: (i) the size, (ii) the distribution of website categories or brands, (iii) the variety of the supported languages, and (iv) the type of elements or resources in the dataset. The rational of each selected factor is discussed in the following subsection.

4.1. The Size

According to the statistical understanding, standard error (SE) is the estimated standard deviation of a parameter, the value of which is not known exactly. The standard error, SE is defined as follows:

$$SE = \frac{\delta}{\sqrt{n}}, \qquad (1)$$

where δ is the sample standard deviation, and n is the size of sample. From Equation (1), we know that when the sample size n is increased, the standard error, SE will decrease. In other words, a larger sample size will broaden the coverage of data and approximate more closely to the real population. Hence, the size of the dataset appears to be one of the important factors in the experiment and needs to be significantly large.

What is the suitable size that can be considered as significantly large for a dataset to be used in the anti-phishing research? For this question, we propose to use the expected value, E from all the datasets listed in Table 4. The expected value, E is referred to the measure of the central tendency of a probability distribution or of the random variables characterised by that distribution [53]. The expected value, E is defined as follows:

$$E = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \qquad (2)$$

where n is the total number of datasets, and x_i is the size of i-th dataset. Based on Equation (2) and ignoring the extreme size of the dataset in Ref. [54], the expected value for the datasets listed in Table 4 is 28,521. Hence, we propose to construct the standard offline dataset with 30,000 samples in this paper. The dataset will contain 15,000 samples for phishing and legitimate webpages, respectively.

4.2. Distribution of Website Categories

In addition, the dataset should cover a wide range of webpage categories. Especially for the phishing dataset, we suggest to use the reports released by Anti-Phishing Working Group (APWG) as the reference for the webpage categorisation and the expect ratio of distribution. It is because the APWG is a worldwide coalition that unifying the global response on cybercrime across different fields [55]. As for the legitimate dataset, we only refer to the top ranking lists (e.g., Alexa top 1 million ranking list) and emphasise less on the categorisation and the distribution ratio of the webpages. This is practical because phishers usually would want to target on website that has higher popularity (i.e., larger user base).

We have collected and summarised in Table 1 the APWG phishing attacks reports of the fourth quarters of the year 2012 [56], 2013 [57], and 2014 [58]. The first column shows the categories of website and their distributions for year 2012 – 2014 are shown in the second to the fourth column, respectively. The fifth column shows the average of the three statistics collected for each category. Through the average of each category, we propose the new distribution ratio for our phishing dataset in the last column.

4.3. Variety of Languages

Another important factor is the variety of the supported languages in the dataset. According to the APWG reports, phishers usually will mount attacks in different countries [56, 57, 58]. Therefore, it is important for a dataset to include different languages. We propose the dataset to include some common world languages, such as English, French, Japanese, Chinese, Hindi, Korean, Arabic and Russian. These languages are all listed as the world most common languages or the top Internet languages [59, 60].

4.4. Various Types of Resources

The content of a sample is the most important factor to be concerned in a good dataset construction. In other words, what kind of information or element is worth keeping in the dataset. A good dataset should have the heterogeneousness to support in multiple feature extraction process in an experiment. In order to find out the potential elements, we study the features used in each publication listed in Table 4, and summarised the findings in Table 2.

From Table 2, the findings show that the important elements to be included in a dataset are: URL, HyperText Markup Language (HTML) files, Cascading Style Sheet (CSS) files, Favicon, screenshot of webpage, and WHOIS information of webpage server. Moreover, we also propose to include some other elements which are necessary to render a complete webpage, such as webpage image resources, Scalable Vector Graphics (SVG) files, JavaScript files, web font files, and other necessary but uncommon file type. Table 3 summarises these elements.

5. Dataset Construction

In this section, we will explain in detail on the process of constructing the proposed offline dataset. The aim is to construct a dataset that contains 30,000 webpages (i.e., 15,000 legitimate webpages and 15,000 phishing webpages) according to the criteria discussed in Section 4. Figure 1 shows the framework of the proposed dataset construction.

Table 1: Distribution of the most targeted categories in phishing attacks

Category	2012 4Q (%)	2013 4Q (%)	2014 4Q (%)	Average (%)	Proposed (%)
Financial	34.40	24.26	20.79	26.48	25.00
Payment Service	32.10	53.95	25.13	37.06	35.00
Gaming	14.70	1.35	1.20	5.75	5.00
ISP	9.50	4.87	2.75	5.71	5.00
Social Networking	6.00	0.54	6.43	4.32	4.00
Retail / Service	5.12	7.79	29.37	14.09	10.00
Auction	2.07	2.43	-	2.25	3.00
Government	1.00	1.39	0.56	0.98	1.00
Classified	0.30	0.43	0.07	0.26	1.00
Email	-	-	12.39	12.39	8.00
Other	6.78	2.99	1.31	3.69	3.00

Table 2: Important elements used in different anti-phishing works

Publication	Main Method	Important Element
[12], [13], [14]	List-based	• URL of the webpage
[15], [16], [18], [19], [20], [21], [22]	Content-based	• HTML file • CSS file
[25], [26], [28], [29], [30], [31], [32]	URL-based	• Webpage URL • Hyperlinks inside the webpage • WHOIS information of webpage server
[41], [42], [44], [45], [46], [47], [48]	Visual-based	• Favicon • Screenshot of web-page • HTML file

Table 3: Elements and the corresponding file extension

No	Element	File Extension
1	URL	.txt
2	HTML files	.htm/.html
3	CSS files	.css
4	Favicon	.ico
5	Image resources	.bmp/.gif/.jpg/.jpeg/.png
6	SVG files	.svg
7	JavaScript files	.js
8	Web font files	.eot/.ttf
9	Screenshot of webpage	.bmp/.gif/.jpg/.jpeg/.png
10	WHOIS information	.txt
11	Other uncommon file type	-

5.1. Data Source Selection

We have selected Alexa, DMOZ, and BOTW [61] as the main sources to construct the legitimate samples of the dataset. Alexa is a commercial website which provides traffic data, global rankings and other information on millions of websites [51]. Similar to Alexa, DMOZ is a multilingual open-content directory of World Wide Web links [52], while BOTW Directory is a commercial web directory that provides websites in different topical and regional categorisation [61]. As mentioned in the previous section, low popularity websites are as important as popular websites. Therefore, we have included BOTW and DMOZ to obtain the low popularity websites. During the data crawling process, we notice that BOTW and DMOZ were able to provide the URLs, which were not included in the Alexa top one million websites. We choose 14,500 URLs from Alexa top one million list to serve as the high popularity websites sample, while 500 URLs are from DMOZ and BOTW to serve as the low popularity websites sample. For the phishing dataset, we decided to utilise PhishTank as our phishing source, as it is by far the most complete repository. We choose 15,000 URLs from PhishTank under the category of 'valid phishes' and 'online' to serve as our phishing crawler input. The 'valid phishes' status refers to the reported website, which has been truly verified as phishing website, while 'online' status means the particular reported website is still live and accessible.

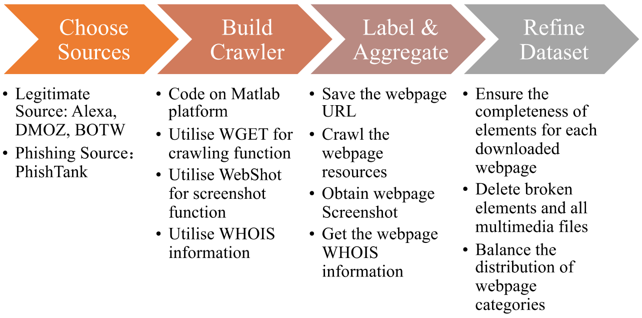
Choose Sources
Build Crawler
Label & Aggregate
Refine Dataset
• Legitimate Source: Alexa, DMOZ, BOTW
• Phishing Source: PhishTank
• Code on Matlab platform
• Utilise WGET for crawling function
• Utilise WebShot for screenshot function
• Utilise WHOIS information
• Save the webpage URL
• Crawl the webpage resources
• Obtain webpage Screenshot
• Get the webpage WHOIS information
• Ensure the completeness of elements for each downloaded webpage
• Delete broken elements and all multimedia files
• Balance the distribution of webpage categories

Fig. 1: The framework of proposed dataset construction

5.2. Webpage Crawler Preparation

We implemented the automated webpage wrapper program using Matlab 2015b and executed on a Windows 10 computer with Intel E3-1230v3 processor and 8GB RAM. The program will read the URLs input from a text file and download the webpages automatically. The wrapper utilised WGET [62] as the crawler function to save the webpages. WGET is a free software package for retrieving files using HTTP, HTTPS and FTP. Besides that, the wrapper also utilised WebShot [63] to take screenshots of webpages and save them as a full-sized images or thumbnails. Furthermore, we utilised WHOIS [64] to obtain the registration record of the domain name and IP address.

5.3. Dataset Labelling and Aggregation

The downloading process started from 20 March 2016 until 30 April 2016. The crawler program will download all the resources (i.e., those resources listed in Table 3) of a rendered webpage. All the downloaded resources will be saved in the designated folder, namely with the folder name of 'L00001' to 'L15000' for each legitimate webpage and 'P00001' to 'P15000' for each phishing webpage. Each designated folder contains 5 subfolders, namely 'RAW-HTML', 'SCREEN-SHOT', 'URL', 'WEBPAGE' and 'WHOIS'.

5.4. Dataset Refinement

The downloaded webpages will go through a manual inspection process. The existence of all subfolders and their contents in each webpage folder will be verified and checked. All multimedia files like music and video (mp3, mp4, avi, wmv, wav, mid, etc.), and files with over-long file name will be excluded. The files with over-long file name are always come in unknown file type, and always referred as the broken element, as it cannot be open, rename, or even deleted. Furthermore, according to the best of our knowledge, there is no evidence in the literature that video or music files are utilised for anti-phishing detection.

As discussed in Section 4, the phishing webpages categorisation is based on the most-targeted sector list in the APWG report. After all the phishing samples have been categorised and organised according to the corresponding categories, we will compute their distribution. Based on the distribution, we will trim a category with excessive samples by removing the duplicate samples. Duplicate samples are those webpages that only differ in the URL. On the contrary, downloading process for the new webpages will be initiated to fill up the undersized category.

Figure 2 summarise the distribution of the proposed dataset. The complete dataset is available for download at http://www.fcsit.unimas.my/research/legit-phish-set

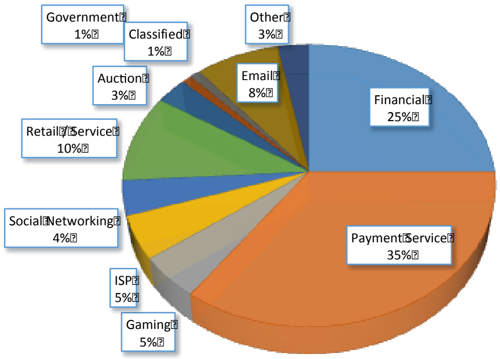

Fig. 2: Distribution of different categories in phishing dataset

6. Conclusion and Future Works

This paper has pointed out the demand for the offline dataset in anti-phishing research community and has constructed an offline dataset. The contributions of this paper include: providing an offline dataset for rapid preliminary method testing and, serving as a permanent repository for phishing webpages. Since, phishing websites have short life span, accessing a shutdown phishing website to retrieve some unique phishing characteristic becomes possible through this dataset.

This paper has reviewed and summarised various datasets from different publications according to a few aspects: dataset size, sources, and downloadable. This paper also discusses on the influencing factors during the dataset construction, which include: the dataset size, the distribution of webpage categories and brands, the variety of the supported languages, and the contents of the dataset.

As for future work, we plan to include samples with different features, such as HTML5 webpage. As the HTML5 becomes more common in the web designs, it is reasonable to include HTML5 webpages to the dataset. This is important because the HTML structure in HTML5 website is different than the previous standard.

Acknowledgement

The funding for this project was made possible through a research grant obtained from UNIMAS under the Special FRGS 2016 Cycle [Grant No: F08/SpFRGS/1533/2017].

References

[1] D. Goel and A. K. Jain, “Mobile phishing attacks and defence mechanisms: State of art and open research challenges,” Computers & Security, vol. 73, pp. 519–544, 2018.

[2] K. L. Chiew, E. H. Chang, S. S. Nah, and W. K. Tiong, “Utilisation of Website Logo for Phishing Detection,” Computer & Security, vol. 54, pp. 16–26, 2015.

[3] G. Ramesh, J. Gupta, and P. Gamya, “Identification of phishing webpages and its target domains by analyzing the feign relationship,” Journal of Information Security and Applications, vol. 35, pp. 75–84, 2017.

[4] A. Aleroud and L. Zhou, “Phishing environments, techniques, and countermeasures: A survey,” Computers & Security, vol. 68, pp. 160–196, 2017.

[5] E. H. Chang, K. L. Chiew, S. S. Nah, and W. K. Tiong, “Phishing Detection via Identification of Website Identity,” 2013 International Conference on IT Convergence and Security, pp. 1–4, 2013.

[6] J. Zhang, Y. Ou, D. Li, and Y. Xin, “A Prior-based Transfer Learning Method for the Phishing Detection,” Journal of Networks, vol. 7, no. 8, pp. 1201–1207, 2012.

[7] I. Kang, P. Kim, S. Lee, H. Jung, and B. You, “Construction of a large-scale test set for author disambiguation,” Information Processing & Management, vol. 47, no. 3, pp. 452–465, 2011.

[8] K. Yuan, Z. Tian, J. Zou, Y. Bai, and Q. You, “Brain CT image database building for computer-aided diagnosis using content-based image retrieval,” Information Processing & Management, vol. 47, no. 2, pp. 176–185, 2011.

[9] P. Bailey, N. Craswell, and D. Hawking, “Engineering a multi-purpose test collection for web retrieval experiments,” Information Processing & Management, vol. 39, no. 6, pp. 853–871, 2003.

[10] J. Zhang, S. Luo, Z. Gong, and Y. Xin, “Protection Against Phishing Attacks: A Survey,” International Journal of Advancements in Computing Technology, vol. 3, no. 9, pp. 155–164, 2011.

[11] F. Schneider, N. Provos, R. Moll, M. Chew, and B. Rakowski, “Phishing protection: Design documentation,” [Online]. Available: https://wiki.mozilla.org/Phishing Protection: Design Documentation, accessed: 23 February 2017.

[12] R. Abrams, O. Barrera, and J. Pathak, “Browser security comparative analysis - phishing protection,” [Online]. Available: https://www.nsslabs.com/index.cfm/_api/render/file/?method=inline&fileID=A02950BF-5056-9046-93D93A5D61314F1D, accessed: 23 February 2017.

[13] Y. Cao, W. Han, and Y. Le, “Anti-phishing Based on Automated Individual White-list,” Proceedings of the 4th Workshop on Digital Identity Management, pp. 51–60, 2008.

[14] P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, “PhishNet: Predictive Blacklisting to Detect Phishing Attacks,” 29th IEEE International Conference on Computer Communications, pp. 346–350, 2010.

[15] Y. Zhang, J. I. Hong, and L. F. Cranor, “Cantina: A Content-based Approach to Detecting Phishing WebSites,” Proceedings of the 16th International Conference on World Wide Web, pp. 639–648, 2007.

[16] G. Xiang and J. I. Hong, “A Hybrid Phish Detection Approach by Identity Discovery and Keywords Retrieval,” Proceedings of the 18th International Conference on World Wide Web, pp. 571–580, 2009.

[17] Q. Cui, G.-V. Jourdan, G. V. Bochmann, R. Couturier, and I.-V. Onut, “Tracking phishing attacks over time,” in Proceedings of the 26th International Conference on World Wide Web, pp. 667–676, 2017.

[18] L. Wenyin, N. Fang, X. Quan, B. Qiu, and G. Liu, “Discovering Phishing Target Based on Semantic Link Network,” Future Generation Computer Systems, pp. 381–388, 2010.

[19] J. Zhang, C. Wu, H. Guan, Q. Wang, L. Zhang, Y. Ou, Y. Xin, and L. Chen, “An Content-analysis Based Large Scale Anti-Phishing Gateway,” 2010 IEEE 12th International Conference on Communication Technology, pp. 979–982, 2010.

[20] C. Soman, H. Pathak, V. Shah, A. Padhye, and A. Inamdar, “An Intelligent System for Phish Detection, using Dynamic Analysis and Template Matching,” International Journal of Computer, Electrical, Automation, Control and Information Engineering, vol. 2, no. 6, pp. 1927–1933, 2008.

[21] R. M. Mohammad, F. A. Thabtah, and L. McCluskey, “Intelligent Rule-based Phishing Websites Classification,” IET Information Security, vol. 8, no. 3, pp. 153–160, 2014.

[22] N. Abdelhamid, “Multi-label rules for phishing classification,” Applied Computing and Informatics, vol. 11, no. 1, pp.29–46, 2015.

[23] A. K. Jain and B. B. Gupta, “A novel approach to protect against phishing attacks at client side using auto-updated white-list,” EURASIP Journal on Information Security, vol. 2016, no. 9, pp. 1–11, May 2016.

[24] R. S. Rao and A. R. Pais, “Detecting phishing websites using automation of human behavior,” in Proceedings of the 3rd ACM Workshop on Cyber-Physical System Security, pp. 33–42, 2017.

[25] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, “Beyond blacklists: learning to detect malicious websites from suspicious URLs,” 15th ACM International Conference on Knowledge Discovery and Data Mining, pp. 1245–1254, 2009.

[26] S. Garera, N. Provos, M. Chew, and A. D. Rubin, “A Framework for Detection and Measurement of Phishing Attacks,” Proceedings of the 2007 ACM Workshop on Recurring Malcode, pp. 1–8, 2007.

[27] Z. Hu, R. Chiong, I. Pranata, W. Susilo, and Y. Bao, “Identifying malicious web domains using machine learning techniques with online credibility and performance data,” in IEEE Congress on Evolutionary Computation, pp. 5186–5194, 2016.

[28] M. Maurer and L. Höfer, “Sophisticated Phishers Make More Spelling Mistakes: Using URL Similarity against Phishing,” 4th International Symposium in Cyberspace Safety and Security, pp. 414–426, 2012.

[29] B. E. Sananse and T. K. Sarode, “Phishing URL Detection: A Machine Learning and Web Mining-based Approach,” International Journal of Computer Applications, vol. 123, no. 13, pp. 46–50, 2015.

[30] L. A. T. Nguyen, B. L. To, H. K. Nguyen, and M. H. Nguyen, “Detecting phishing web sites: A heuristic URL-based approach,” 2013 International Conference on Advanced Technologies for Communications, pp. 597–602, 2013.

[31] R. Verma and K. Dyer, “On the Character of Phishing URLs: Accurate and Robust Statistical Learning Classifiers,” Proceedings of the 5th ACM Conference on Data and Application Security and Privacy, pp. 111–122, 2015.

[32] W. Chu, B. B. Zhu, F. Xue, X. Guan, and Z. Cai, “Protect sensitive sites from phishing attacks using features extractable from inaccessible phishing URLs,” IEEE International Conference on Communications, pp. 1990–1994, 2013.

[33] A. Y. Daeef, R. B. Ahmad, Y. Yacob, and N. Y. Phing, “Wide scope and fast websites phishing detection using URLs lexical features,” in 3rd International Conference on Electronic Design, pp. 410–415, 2016.

[34] J. Solanki and R. G. Vaishnav, “Website phishing detection using heuristic based approach,” International Research Journal of Engineering and Technology, vol. 3, no. 5, pp. 2044–2048, May 2016.

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

[36] Z. Yan, S. Liu, T. Wang, B. Sun, H. Jiang, and H. Yang, “A genetic algorithm based model for chinese phishing e-commerce websites detection,” in HCI in Business, Government, and Organizations: eCommerce and Innovation, pp. 270–279, 2016.

[37] S. Marchal, G. Armano, T. Gröndahl, K. Saari, N. Singh, and N. Asokan, “Off-the-hook: An Efficient and Usable Client-Side Phishing Prevention Application,” IEEE Transactions on Computers, vol. 66, no. 10, pp. 1717–1733, 2017.

[38] R. Jabri and B. Ibrahim, “Phishing websites detection using data mining classification model,” Transactions on Machine Learning and Artificial Intelligence, vol. 3, no. 4, pp. 42–51, Sep. 2015.

[39] R. Gupta and P. K. Shukla, “Performance analysis of anti-phishing tools and study of classification data mining algorithms for a novel anti-phishing system,” International Journal of Computer Network and Information Security, vol. 12, pp. 70–77, Nov. 2015.

[40] P. Singh, Y. P. Maravi, and S. Sharma, “Phishing websites detection through supervised learning networks,” in International Conference on Computing and Communications Technologies, pp. 61–65, 2015.

[41] W. Liu, X. Deng, G. Huang, and A. Y. Fu, “An Antiphishing Strategy Based on Visual Similarity Assessment,” IEEE Internet Computing, vol. 10, no. 2, pp. 58–65, 2006.

[42] A. Y. Fu, L. Wenyin, and X. Deng, “Detecting Phishing Web Pages with Visual Similarity Assessment Based on Earth Mover’s Distance (EMD),” IEEE Transactions on Dependable and Secure Computing, vol. 3, no. 4, pp. 301–311, 2006.

[43] R. S. Rao and S. T. Ali, “A computer vision technique to detect phishing attacks,” in Fifth International Conference on Communication Systems and Network Technologies, pp. 596–601, 2015.

[44] M. Dunlop, S. Groat, and D. Shelly, “GoldPhish: Using Images for Content-Based Phishing Analysis,” Proceedings of the 2010 Fifth International Conference on Internet Monitoring and Protection, pp. 123–128, 2010.

[45] J. C. S. Fatt, K. L. Chiew, and S. N. Sze, “Phishdentity: Leverage Website Favicon to Offset Polymorphic Phishing Website,” Ninth International Conference on Availability, Reliability and Security, pp. 114–119, 2014.

[46] J. Mao, P. Li, K. Li, T. Wei, and Z. Liang, “BaitAlarm: Detecting Phishing Sites Using Similarity in Fundamental Visual Features,” 5th International Conference on Intelligent Networking and Collaborative Systems, pp. 790–795, 2013.

[47] W. Zhang, H. Lu, B. Xu, and H. Yang, “Web Phishing Detection Based on Page Spatial Layout Similarity,” Informatica (Slovenia), vol. 37, no. 3, pp. 231–244, 2013.

[48] M. Hara, A. Yamada, and Y. Miyake, “Visual similarity-based phishing detection without victim site information,” 2009 IEEE Symposium on Computational Intelligence in Cyber Security, pp. 30–36, 2009.

[49] “Google.com,” [Online]. Available: https://www.google.com/, accessed: 4 April 2017.

[50] “Yahoo.com,” [Online]. Available: https://www.yahoo.com/, accessed: 4 April 2017.

[51] Alexa Internet Inc., “Keyword Research, Competitive Analysis, & Website Ranking,” [Online]. Available: http://www.alexa.com/, accessed: 4 April 2017.

[52] “Dmoz.com,” [Online]. Available: https://www.dmoz.org/, accessed: 4 April 2017.

[53] W. Feller, Introduction to Probability Theory and Its Applications, 3rd ed. Wiley, 1968, vol. 1.

[54] H. Chen, A. Abbasi, B. Thuraisingham, C. Yang, P. Hu, and R. Shenandoah, “Internet phishing websites,” [Online]. Available: http://www.azsecure-data.org/phishing-websites.html, 2017, accessed: 7 June 2017.

[55] APWG, “About the APWG,” [Online]. Available: http://www.antiphishing.org/about-APWG, accessed: 23 February 2017.

[56] APWG, “Phishing Activity Trends Reports 4th Quarter 2012,” [Online]. Available: http://docs.apwg.org/reports/apwg_trends_report_Q4_2012.pdf, accessed: 23 February 2017.

[57] APWG, “Phishing Activity Trends Reports 4th Quarter 2013,” [Online]. Available: http://docs.apwg.org/reports/apwg_trends_report_q4_2013.pdf, accessed: 23 February 2017.

[58] APWG, “Phishing Activity Trends Reports 4th Quarter 2014,” [Online]. Available: http://docs.apwg.org/reports/apwg_trends_report_q4_2014.pdf, accessed: 23 February 2017.

[59] One World Nations Online, “Most widely spoken languages in the world,” [Online]. Available: http://www.nationsonline.org/oneworld/most_spoken_languages.htm, accessed: 4 April 2017.

[60] Internet World Stats, “Internet world users by language,” [Online]. Available: http://www.internetworldstats.com/stats7.htm, accessed: 4 April 2017.

[61] “Best of the Web,” [Online]. Available: http://botw.org, accessed: 4 April 2017.

[62] “GNU Wget,” [Online]. Available: https://www.gnu.org/software/wget/, accessed: 4 April 2017.

[63] “WebShot,” [Online]. Available: http://www.websitescreenshots.com/, accessed: 4 April 2017.

[64] Microsoft Corporation, “Whois - Windows Sysinternals,” [Online]. Available: https://docs.microsoft.com/en-us/sysinternals/downloads/whois/, accessed:10 March 2018.

[65] R. M. A. Mohammad, L. McCluskey, and F. Thabtah, “Phishing websites data set,” [Online]. Available: https://archive.ics.uci.edu/ml/datasets/phishing+websites, accessed: 7 June 2017.

[66] N. Abdelhamid, “Website phishing data set,” [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Website+Phishing, 2016, accessed: 7 June 2017.

Table 4: Datasets information

Publication	Dataset Size	Source of the Dataset	Downloadable
[12]	Total: 168,462	-	No
[13]	Training: Legit: 37 Phish: 28	• www.phishtank.com • email • blog post	No
[14]	Legit: 120,000 Phish: 32,000	• www.phishtank.com • spamscatter.com • www.cutestat.com • www.dmoz.org • YRUG (Yahoo Random URL Generator)	No
[15]	Legit: 100 Phish: 100	• www.phishtank.com • www.3sharp.com	No
[16]	Legit: 3,543 Phish: 7,906	• www.google.com • www.millersmiles.co.uk • uribl.com • ww.phishtank.com	No
[17]	Legit: 24,800 Phish: 19,066	• www.phishtank.com • www.alexa.com	Yes
[18]	Legit: 1,000 Phish: 1,000	• www.phishtank.com	No
[19]	Legit: 93,000 Phish: 123,521	• www.phishtank.com • www.aa419.org • apwg.org	No
[20]	Total: 70-90	-	No
[21]	Legit: 450 Phish: 2,500	• www.phishtank.com • www.millersmiles.co.uk • www.yahoo.com • www.stpt.com/directory	No
[22]	Legit: 548 Phish: 805	• www.phishtank.com • www.millersmiles.co.uk • www.yahoo.com • www.stpt.com/directory	No
[23]	Legit: 405 Phish: 1,120	• www.phishtank.com • www.alexa.com • stuffgate.com • Online payment service provider	No
[24]	Legit: 883 Phish: 1,459	• www.phishtank.com • www.alexa.com	No
[25]	Legit: 15,000 Phish: 20,500	• www.yahoo.com • www.dmoz.org • www.phishtank.com • spamscatter.com.cutestat.com	No
[26]	Legit: 1,263 Phish: 1,245	• Google Toolbar URL (white and black list)	No
[27]	Legit: 1,000 Phish: 1,000	• www.phishtank.com • www.alexa.com	No
[28]	Legit: 127 Phish: 8,730 (only 566 have screenshot)	• www.phishtank.com	No
[29]	Total: 600	• www.phishtank.com • www.google.com	No
[30]	Legit: 1,000 Phish: 10,661	• www.phishtank.com • www.dmoz.org	No
[31]	Dataset 1 Legit: 13,274 Phish: 11,271 Dataset 2 Legit: 14,999 Phish: 14,920 Dataset 3 Legit: 19,999 Phish: 18,395 Dataset 4 Legit: 11,275 Phish: 11,271	• www.alexa.com • www.dmoz.org • www.phishtank.com • apwg.org	No
[32]	Legit: 28,722 Phish: 17,423	• www.taobao.com • www.yahoo.com • www.hao123.com • www.baidu.com	No
[33]	Legit: 20,550 Phish: 54,287	• www.phishtank.com • openphish.com • dmoz.org • webcrawler.com	No
[34]	Total: 200	• www.phishtank.com • Yahoo	No

[36]	Legit: 1,462 Phish: 1,416	• www.315online.com.cn • www.anquan.org	No
[37]	Legit: 208,500 Phish: 16,409	• www.phishtank.com • www.alexa.com	No (on request)
[38]	Legit: 300 Phish: 700	• www.phishtank.com	No
[39]	Legit: 8,540 Phish: 4,480	• www.phishtank.com • APWG database	No
[40]	Legit: 179 Phish: 179	• www.phishtank.com • www.alexa.com	No
[41]	Legit: 320 Phish: 8	• apwg.org	Yes
[42]	Legit: 10,272 Phish: 9	• www.google.com	Yes
[43]	Legit: 20 Phish: 400	• www.phishtank.com	No
[44]	Legit: 100 Phish: 100	• www.phishtank.com • www.randomwebsite.com • www.web100.org	No
[45]	Legit: 500 Phish: 500	• www.alexa.org • www.phishtank.com	No
[46]	Phish: 7,764	• www.phishtank.com	No
[47]	Legit: 100 Phish: 100	• www.phishtank.com • www.yahoo.org	No
[48]	Legit: 521 Phish: 2,262	• www.phishtank.com • www.alexa.com	No
[65]	Total: 2,456	• www.phishtank.com • MillerSmiles • Google	Yes
[66]	Legit: 548 Phish: 702 Suspicious: 103	• www.phishtank.com • Yahoo	Yes
[54]	Legit: 200 Phish: ~1.8 millions	• www.phishtank.com • Escrow Fraud Prevention • www.legitscript.com • PhishMonger	Yes