Phishing Webpage Detection Using Weighted URL Tokens for Identity Keywords Retrieval

Choon Lin Tan, Kang Leng Chiew and San Nah Sze

Abstract Phishing is an online identity theft that has threatened Internet users for more than a decade. This paper proposes an anti-phishing technique based on a weighted URL tokens system, which extracts identity keywords from a query webpage. Using the identity keywords as search terms, a search engine is invoked to pinpoint the target domain name, which can be used to determine the legitimacy of the query webpage. Experiments were conducted over 1000 datasets, where 99.20 % true positives and 92.20 % true negatives were achieved. Results suggest that the proposed system can detect phishing webpages effectively without using conventional language-dependent keywords extraction algorithms.

Keywords Phishing detection · Identity keywords · Keywords retrieval · Search engine · Weighted URL tokens

1 Introduction

Phishing websites are counterfeit websites designed to deceive victims in order to steal their sensitive information. Phishers usually entice victims to the phishing website by sending emails containing the fraudulent URL and some threatening messages (e.g., account termination, data loss, etc.). At the phishing website, the phishers will capture every information submitted by the victims.

C.L. Tan (✉) · K.L. Chiew · S.N. Sze
Faculty of Computer Science and Information Technology,
Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
e-mail: colin89lin@gmail.com

K.L. Chiew
e-mail: klchiew@unimas.my

S.N. Sze
e-mail: snsze@unimas.my

H. Ibrahim et al. (eds.), *9th International Conference on Robotic, Vision, Signal Processing and Power Applications*, Lecture Notes in Electrical Engineering 398,
DOI 10.1007/978-981-10-1721-6_15

The severity of phishing threats is on the rise. For instance, a total of 42,212 unique phishing websites was reported in June 2014, as compared to 38,110 in June 2013 [1, 2]. The RSA monthly fraud report [3] stated an estimated loss of $453 million faced by worldwide organization in December 2014.

To protect users from being phished, a novel technique called the weighted URL tokens system is proposed, which finds the target identity to be compared against the actual identity of the query webpage. Here, target identity is defined as the domain name belonging to a legitimate brand that the phishing webpage deceptively represents while actual identity refers to the query webpage's domain name. For legitimate webpages, the target identity often point to its own domain name, while phishing webpage does not. As such, the query webpage can be considered as phishing when its actual domain name fails to match the target domain name. Hereinafter, the term "identity" and "domain name" shall be used interchangeably.

The remainder of this paper is organized as follow: Sect. 2 briefly reviews a number of related works. Section 3 presents the proposed method. Section 4 discusses the experiment setup and results. Finally, Sect. 5 concludes the paper and provides some insight to our future work.

2 Related Works

This section briefly reviews three major categories of conventional anti-phishing techniques that have been introduced over the years.

Zhang et al. [4] propose CANTINA, a text-based anti-phishing technique that extracts keywords from a webpage using the term frequency-inverse document frequency (TF-IDF) algorithm. The keywords are searched on Google to check whether the query webpage domain name exists among the search results. TF-IDF relies on language-specific word list, thus CANTINA is only effective in classifying English webpages. Similar language limitation is found in [5, 6].

Wenyin et al. [7] focus on webpage identity analysis and propose the Semantic Link Network (SLN) consisting of weighted paths linking a set of webpages associated with the query webpage. Several metrics of the SLN are calculated to find the target identity. A similar strategy is proposed in [6]. However, these systems can be bypassed by phishing webpages that contain no legitimate hyperlinks.

Fu et al. [8] assess the visual similarity between the suspected webpages and legitimate webpages using the Earth Movers Distance (EMD). However, phishers can evade their method by altering the position and layout of the visual elements. This weakness is addressed in [9] using the Scale Invariant Feature Transform (SIFT), which are robust against image resizing, rotation, and distortion.

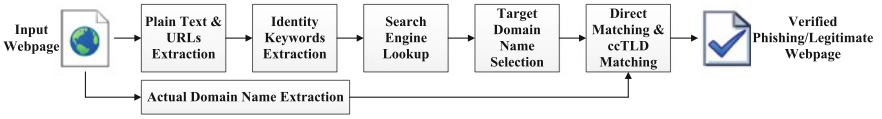

Fig. 1 The architecture of proposed method

3 Methodology

The proposed method focuses on client-side protection, specifically at the instance when the user arrives at a phishing webpage. Figure 1 shows the components of the proposed method, which will be discussed in the following subsections.

3.1 Plain Text and URLs Extraction

Plain texts are extracted from several identity-relevant tags in the HTML source code such as the meta tags, title tag, body tag and *alt* attribute of all tags. On the other hand, URLs are retrieved from the *src* and *href* attributes of tags within the HTML source code.

3.2 Weighted URL Tokens for Identity Keywords Retrieval

Conventional approaches mostly employ the TF-IDF algorithm to extract keywords. To overcome the limitation of TF-IDF discussed in Sect. 2, we introduce a novel identity keywords retrieval system using weighted URL tokens that serves as a robust weight generator.

Figure 2 shows the common appearance of a URL in the web browser. It is evident that words appearing nearer to the left hand side (LHS) of the URL are more likely to capture the attention of users. Phishers exploit this visual appearance by intentionally placing identity keywords towards the LHS of the URL to mimic the appearance of legitimate URLs. Based on this, a word is assumed to be more important when it appears on the LHS. This forms the basis of the proposed weighted URL tokens system.

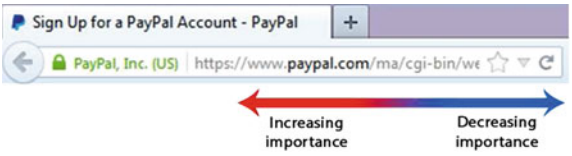

Fig. 2 Perception of users when looking at URL in web browser

Table 1 Level of URL tokens

Level	URL tokens, T_{url}
-	https:
1	//www.paypal.com
2	/ma
3	/cgi-bin
4	/webscr?cmd=_registration-run&from=PayPal

Let the extracted tokens of the plain texts and URL be denoted as T_{plain} and T_{url}, respectively. If T_{plain} is a substring in T_{url}, a weight will be assigned to T_{plain}. The process begins by segmenting the URLs into tokens delimited by the forward slash. Taking the URL https://www.paypal.com/ma/cgi-bin/webscr?cmd=_registration-run&from=PayPal as an example, the segmentation results are shown in Table 1.

Next, the importance level shown in Fig. 2 is quantified. Given the i-th distinct word in the HTML textual content, its final weight, denoted by W_i, is calculated as Eq. 1.

$$W_i = \frac{l_i}{n} \sum_{k=1}^{L} \frac{N_k}{k^2} \quad . \tag{1}$$

where l_i denotes the length of the i-th distinct word, n represents the total number of URLs extracted from the webpage, k is the level of URL where the i-th word occurs, L is the total number of levels available, and N_k is the total number of occurrence of the i-th distinct word in level k. Using Eq. 1, greater weights can be assigned to identity keywords. The top 5 words are selected and filtered to eliminate words with significantly lower scores. The final identity keywords will consists of either 1 or 2 keywords.

3.3 Search Engine Lookup and Target Domain Name Selection

In this stage, the identity keywords are searched on Yahoo search engine to gather the top 30 search results. To select the target domain name from among the search results, the same concept of weighted URL tokens is applied. Instead of calculating weight for T_{plain}, the weighted URL tokens system switches to calculate weight for each domain name in the search results and select the domain name with the highest weight.

3.4 Direct Domain Name Matching and ccTLD Matching

With the target domain name and actual domain name in hand, direct string matching is performed. If both domain names are identical, the query webpage is considered legitimate. Otherwise, the processing continues to the country code Top Level Domain (ccTLD) matching.

The ccTLD consist of only two letters which represent the abbreviation of different countries or territories, such as `.uk`, `.au`, `.fr`, etc. For a particular online service, it might be accessible through different URL such as https://www.amazon.com or https://www.amazon.co.uk. Since both websites belong to Amazon, their identity should be considered the same. Therefore, the module in this section matches the ccTLD using the Root Zone Database[1] of the Internet Assigned Numbers Authority (IANA). If the ccTLD matches, the query webpage is considered legitimate. Otherwise, it is concluded as phishing.

4 Results and Discussion

In this section, the experimental results of the proposed phishing detection method are presented and compared against the results achieved by CANTINA [4] using the same dataset. Specifically, 500 phishing webpages are collected from PhishTank[2] and another 500 legitimate webpages are collected based on the top one million list from Alexa.[3] Before conducting the experiments, the dataset is filtered by manually removing duplicate webpages and webpages that fail to load properly. For benchmarking purposes, the proposed method and CANTINA were prototyped using the Python programming language.

Table 2 shows the evaluation results on 271 samples of non-English webpages within the dataset. Results suggest that the proposed method outperforms CANTINA, scoring an overall accuracy of 91.51 %. This implies that the proposed method is effective in classifying non-English webpages.

In another evaluation, the non-English and English webpages are combined and tested. The overall results are shown in Table 3. It is observed that the proposed method can effectively detect more phishing webpages compared to CANTINA, achieving a true positive rate of 99.20 %. This type of key advantage is very desirable in anti-phishing systems. As for legitimate webpages classification, the proposed method also proved to be superior over CANTINA, scoring a true negative rate of 92.20 %. In summary, the proposed method is capable of classifying both English and non-English webpages effectively, achieving an overall accuracy of 95.70 %.

[1] http://www.iana.org/domains/root/db/.

[2] http://www.phishtank.com/.

[3] http://www.alexa.com/.

Table 2 Comparison of detection performance on non-English webpages

Anti-phishing methods	True positive rate (%)	False negative rate (%)	True negative rate (%)	False positive rate (%)	Overall accuracy (%)
Proposed method	100.00	0.00	88.38	11.62	91.51
CANTINA [4]	91.78	8.20	80.81	19.19	83.76

Table 3 Comparison of overall detection performance

Anti-phishing methods	True positive rate (%)	False negative rate (%)	True negative rate (%)	False positive rate (%)	Overall accuracy (%)
Proposed method	99.20	0.80	92.20	7.80	95.70
CANTINA [4]	94.60	5.40	88.80	11.20	91.70

4.1 System Limitations

This subsection discusses some limitations in the proposed method. First, when a webpage uses images to replace the textual content of the whole webpage, the proposed method might find insufficient identity keywords. This limitation can be resolved by invoking an optical character recognition (OCR) module to extract text in the images and feed them to the proposed system. Second, if phishers successfully registers a non-existent ccTLD-based domain name using a second level domain (SLD) that is identical to the legitimate domain name, the proposed system will produce false negative results.

5 Conclusion and Future Work

In this work, an identity keywords retrieval system based on weighted URL tokens is proposed to detect phishing webpages. Using the identity keywords, a search engine is invoked to obtain the target domain name. The legitimacy of the query webpage is determined by comparing its target domain name and actual domain name. Experiments have showed promising results, where 99.20 % true positives and 92.20 % true negatives were achieved.

In future, the proposed method can be extended to extract identity keywords consisting of multiple words and to handle webpages containing non-ASCII characters (e.g., Chinese, Japanese, Russian, etc.).

Acknowledgments The funding for this project is made possible through the research grant obtained from UNIMAS and the Ministry of Education, Malaysia under the Fundamental Research Grant Scheme 2/2013 [Grant No: FRGS/ICT07(01)/1057/2013(03)].

References

1. Anti-Phishing Working Group: Phishing activity trends report, 2nd quarter 2013 (Nov 2013). http://docs.apwg.org/reports/apwg_trends_report_q2_2013.pdf
2. Anti-Phishing Working Group: Phishing activity trends report, 2nd quarter 2014 (Aug 2014). http://docs.apwg.org/reports/apwg_trends_report_q2_2014.pdf
3. EMC Corporation: RSA monthly fraud report (Jan 2015). http://australia.emc.com/collateral/fraud-report/h13929-rsa-fraud-report-jan-2015.pdf
4. Zhang Y, Hong JI, Cranor LF (2007) CANTINA: a content-based approach to detecting phishing web sites. In: Proceedings of the 16th international conference on World Wide Web. ACM, pp 639–648
5. He M, Horng SJ, Fan P, Khan MK, Run RS, Lai JL, Chen RJ, Sutanto A (2011) An efficient phishing webpage detector. Expert Syst Appl 38(10):12018–12027
6. Ramesh G, Krishnamurthi I, Kumar KSS (2014) An efficacious method for detecting phishing webpages through target domain identification. Decis Support Syst 61:12–22
7. Wenyin L, Fang N, Quan X, Qiu B, Liu G (2010) Discovering phishing target based on semantic link network. Future Gener Comput Syst 26(3):381–388
8. Fu AY, Wenyin L, Deng X (2006) Detecting phishing web pages with visual similarity assessment based on Earth Mover's Distance (EMD). IEEE Trans Dependable Secure Comput 3(4):301–311
9. Huang CY, Ma SP, Yeh WL, Lin CY, Liu CT (2010) Mitigate web phishing using site signatures. In: TENCON 2010-2010 IEEE region 10 conference. IEEE, pp 803–808