PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning

Ankit Kumar Jain and B. B. Gupta

Abstract Today, phishing is one of the most serious cyber-security threat in which attackers steal sensitive information such as personal identification number (PIN), credit card details, login, password, etc., from Internet users. In this paper, we proposed a machine learning based anti-phishing system (i.e., named as PHISH-SAFE) based on Uniform Resource Locator (URL) features. To evaluate the performance of our proposed system, we have taken 14 features from URL to detect a website as a phishing or non-phishing. The proposed system is trained using more than 33,000 phishing and legitimate URLs with SVM and Naïve Bayes classifiers. Our experiment results show more than 90% accuracy in detecting phishing websites using SVM classifier.

Keywords Phishing · SVM · Bayes classifier · Machine learning
URL

1 Introduction

Phishing is one of the major security threats faced by the cyber-world and could lead to financial losses for both industries and individuals. In this attack, criminal makes a fake web page by copying contents of the legitimate page, so that a user cannot differentiate between phishing and legitimate sites [1]. Life cycle of phishing attack is shown in Fig. 1. According to anti-phishing working report in the first Quarter of 2014, second highest number of phishing attacks ever recorded between January and March 2014 [2] and payment services are the most targeted by these attacks. The total number of phishing attacks notice in Q1 (first quarter) of

A. K. Jain (✉) · B. B. Gupta
Department of Computer Engineering, National Institute of Technology Kurukshetra, Kurukshetra 136119, Haryana, India
e-mail: ankit.jain2407@gmail.com

B. B. Gupta
e-mail: gupta.brij@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

M. U. Bokhari et al. (eds.), *Cyber Security*, Advances in Intelligent Systems and Computing 729, https://doi.org/10.1007/978-981-10-8536-9_44

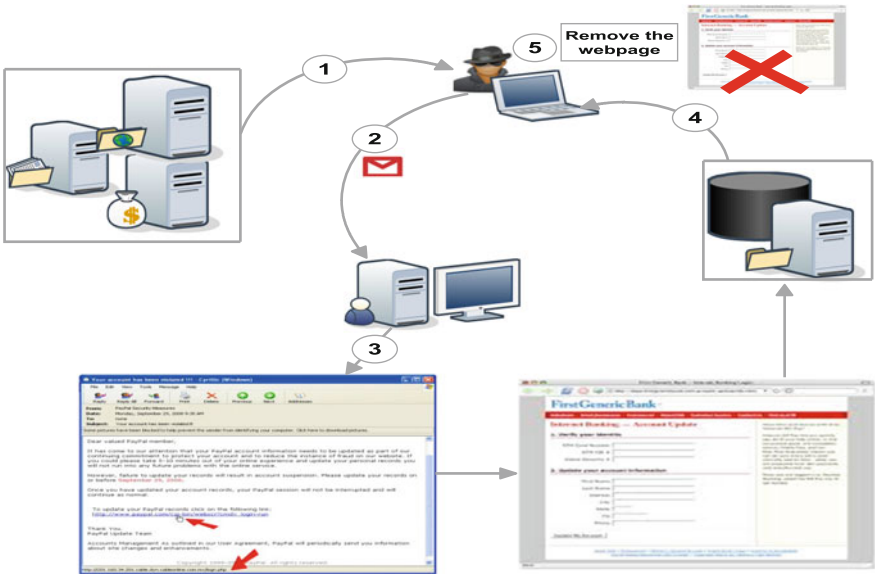

Fig. 1 Phishing life cycle: (1) phisher copies the content from legitimate site and constructs the phishing site; (2) phisher sent link of phishing URL to Internet user; (3) user opens the link and fills personal on fake site; (4) phisher steals the personal information of user; (5) phisher deletes the fake web page

2014 were 125,215, a 10.7 percent increase over Q4 (fourth quarter) of 2013. Existing solution like heuristic based, visual similarity based take features from the web page content so they take a lot of time to take decision. The phishing URL classification scheme based only on investigative the suspicious URL and speed up the running time of system. Therefore, in this paper, we proposed a machine learning based phishing detection system which uses the URL features and analysed it using naive Bayesian and SVM classifiers. Moreover, it does not require any information from the e-content of the suspicious web page.

The remainder of this paper is organized as follows. Section 2 describes the background and state-of-art techniques, its advantages and limitations. Section 3 describes our proposed phishing detection system in details. Evaluation of the proposed system with results is discussed in Sect. 4. Finally, Sect. 5 concludes our paper and discusses the scope for future work.

2 Related Work

There have been several techniques given in the literature to detect phishing attack in last few years. In this section, we present an overview of detection approaches against phishing attacks. Phishing detection approaches are broadly classified into two types: user education based techniques and software-based techniques.

Software-based detection is further classified into heuristic based, blacklist-based and visual similarity based techniques.

User Education based approaches: To classify phishing and non-phishing email, Kumaraguru et al. [3] developed two embedded training designs to teach users. After this training, users can identify phishing emails by themselves. Sheng et al. [4] proposed an educational interactive game "Anti-Phishing Phill" that educates good habits to keep away from phishing attacks.

Softwar-based approaches: Software-based detection is further classified into following sub-categories:

(a) Blackl-based approaches: In this type of approaches, the suspicious domain is matched with a predefined phishing domain called blacklist. The negative aspect of this scheme is that it usually does not cover all phishing websites because a freshly launched fraud website takes some time to add to the blacklist record. Sheng et al. [5] depicted that blacklists are typically add in the record at diverse frequencies, approximate 50–80% of phishing domains added in blacklist after performing some financial loss.
(b) Heuris-based approaches: In this type of approaches, the heuristic design of suspicious websites matches with the feature set, which are generally found in phishing websites [6]. Zero-day attack (i.e., attacks that were not seen before) can be identified using heuristic approach. Zhang et al. [7] proposed a content-based phishing detection technique called CANTINA, which take a rich set of feature set from various field of a web page.
(c) Visual similarity-based approaches: Visual similarity-based approaches compare the visual appearance of a suspicious website and its corresponding legitimate site. Visual similarity-based techniques use features set like text content, HTML Tags, Cascading Style Sheet (CSS), image processing, etc., to make decision. Chen et al. [8] proposed an anti-phishing approach based on discriminative key-point features in a web page.

Based on the abovementioned approaches proposed in the literature, we found that there exists no single technique that can detect various types of phishing attacks. Moreover, Blacklist/White-list based approaches cannot detect zero-day attacks. Heuristic-based techniques can detect the zero-day attack but fail to detect attack if embedded object present in the web page and false positive is also high in these approaches. Moreover, visual similarity-based approaches can detect the embedded objects present in the web page but they fail to detect the zero-day attacks. Therefore, in this paper, we have proposed a machine learning based anti-phishing system (i.e., named as PHISH-SAFE) based on Uniform Resource Locator (URL) features which can able to detect variety of phishing attacks efficiently.

3 Proposed Phishing Detection System

In this section, we will discuss our proposed phishing detection system which can detect a phishing page before user inputs personal information. Total 32,951 phishing URLs are taken from phishtank.com to evaluate the performance of the proposed system. Following features are used for the phishing detection:

- **IP Address**: A phisher uses the IP address in place of domain name to hide the identity of a website.
- **Sub Domain**: Phishing sites contain more than two sub-domains in URL. Each domain is separated by dot (.). If any URL contain three or more than three dots, then the probability of the suspicious site is more. In our experiment, we found that 12,904 sites contain three or more number of dots.
- **URL contains "@" symbol**: the presence of "@" symbol in the URL ignore everything previous to it. In our dataset, out of 32,951 phishing URL, 569 sites contain @ symbol.
- **Number of dash (-) in URL**: To looks like genuine URL, phisher adds some prefix or suffix with the brand name with dash, e.g., www.amazon-india.com. We found that 42.5% of phishing URLs contain "dash" symbol.
- **Length of URL**: To hide the domain name, phisher uses the long URL. In our experiment, we found the average length of URL is 74. We found that 7406 phishing sites contain length between 14 and 40 characters. 10,466 phishing sites are having length between 41 and 60 characters. 6602 phishing URL contain length between 61 and 80 character and 8475 sites contain length between 81 and 2205 characters.
- **Suspicious words in URL**: Phishing URLs contain suspicious words such as token, confirm, security, PayPal, login, signin, bank, account, update, etc., to gain the trust on website. We have taken these nine frequently occurred words in phishing sites.
- **Position of Top-Level Domain**: This feature checks the position of top-level domain at proper place in URL.
  Example—http://xyz.paypal.com.accounts.765issapidll.xtmll.ebmdata.com.
- **Embedded Domain in URL**: It checks this by checking for the occurrence of "//" in the URL.
- **HTTPS Protocol**: HTTPS protocol is used for security. Phishing does not start with https while legitimate URL provides security. (In our phishing dataset, only 388 phishing sites contain https protocol).
- **Number of times http appears**: In phishing websites, http protocol may appear more than one time but in genuine site, it appear only one time.
- **Domains count in URL**: Phishing URL may contain more than one domain in URL. Two or more domains is used to redirect address.
- **DNS lookup**: If the DNS record is not available then the website is phishing. The life of phishing site is very short, therefore; this DNS information may not be available after some time.

- **Inconsistent URL**: If the domain name of suspicious web page is not matched with the WHOIS database record, then the web page is considered as phishing.
- **Age of Domain**: If the age of website is less than 6 month, then chances of fake web page are more.

Training and testing of the proposed system are performed using following classifiers:

(a) Naïve Bayes: Naïve Bayes is the probabilistic classifier, based on Bayes' theorem with "naive" independence supposition. This classifier, used in text categorization, can be an earning-based variant of keyword filtering. The rules for decision making are explained below:

$$\emptyset_{k|y=1} = p\left(x_j = k|y = 1\right) = \left(\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} 1\left\{x_j^{(i)} = k \text{and} y^{(i)} = 1\right\} + 1}{\left(\sum_{i=1}^{m} 1\{y^{(I)} = 1\} n_i\right) + |V|}\right) \quad (1)$$

$$\emptyset_{k|y=1} = p\left(x_j = k|y = 0\right) = \left(\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} 1\left\{x_j^{(i)} = k \text{ and } y^{(i)} = 0\right\} + 1}{\left(\sum_{i=1}^{m} 1\{y^{(I)} = 1\} n_i\right) + |V|}\right) \quad (2)$$

$$\emptyset_{y=1} = \frac{\sum_{i=1}^{m} 1\left\{y^{(i)} = 1\right\}}{(m)} \quad (3)$$

$\emptyset_{x|y=1}$ estimates the probability that a particular feature in a phishing URL will be the k-th word in the dictionary. $\emptyset_{x|y=0}$ estimates the probability that a particular feature in a legitimate URL will be the k-th word in the dictionary. $\emptyset_y$ estimates the probability that any particular URL will be a phishing URL. m is the number of URLs in our training set. The entire dictionary contains V words or the entire URLs are V in number. For training, $\emptyset_{x|y=0}$, $\emptyset_{x|y=1}, \emptyset_y$ are calculated and for testing, $p(x|y = 1)\ p(y = 1)$ is compared to $p(x|y = 0)\ p(y = 0)$. To avoid underflow error, logarithms are used. An email is classified as spam or phishing according to the following equation:

$$\log p(x|y = 1) + \log p(y = 1) > \log p(x|y = 0) + \log p(y = 0) \quad (4)$$

Support Vector Machine: Support vector machine (SVM) is supervised learning models frequently used classifier in phishing attack detection. SVM worked based on training examples and a predefined alteration θ:Rs → F, it makes a map from features set to produce a transformed feature space, storing the URL samples of the two classes with a hyperplane in the transformed feature space.

Table 1 Experiment results

URL instances	Classifiers	
	Naive Bayes (%)	SVM (%)
10,000	64.74	76.04
25,000	76.87	91.28

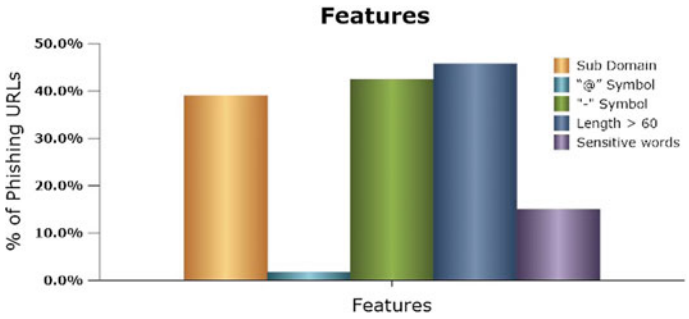

Fig. 2 Features contain by phishing URLs

4 Results and Discussion

In this section, we will discuss the tools and datasets used for implementation and experiments results. The phishing detection using machine learning is classification problem where system learns using various features of phishing and legitimate URLs. After learning the system takes decision automatically based on training. We have recognized various features of phishing and legitimate URLs discussed in the previous section. We have collected 32,951 phishing URLs, taken from PhishTank [9] and 2500 legitimate URLs taken from various sources.

Dataset Used: The dataset for phishing URLs is downloaded from PhishTank. On 20th March 2015, a set of 32,951 phishing URLs were downloaded from PhishTank. The datasets for non-phishing URLs are downloaded from Yahoo Directory by using LinkKlipper from Chrome and DMOZ open directory.

Experiment Results: The feature extraction algorithm is implemented in Java and the features of the URLs are stored in rows of a Sparse Matrix. A set of 15,000 training data (14,000 phishing URLs and 1000 non-phishing URLs) produced an accuracy of 76.04%. A set of 25,000 training URLs (23,000 phishing URLs and 2000 non-phishing URLs) produced an accuracy of 91.28%. Phishing URL detection using Naïve Bayes and SVM classifiers produced the results shown in Table 1. From Table 1, it is found that when the size of the training set increases, SVM performs better than Naïve Bayes classifier to detect phishing URL. Figure 2 shows the features contain by phishing URLs.

5 Conclusion and Future Scope

This paper presented our proposed phishing detection system based on machine learning. We have used 14 different features that distinguish phishing websites from legitimate websites. Our experiment results show more than 90% accuracy in detecting phishing websites using SVM classifier. In future, more features can be added to improve the accuracy of the proposed phishing detection system. Furthermore, other machine learning techniques can be used to increase the efficiency of the proposed system.

References

1. Almomani A, Gupta BB, Atawneh S, Meulenberg A, Almomani E (2013) A survey of phishing email filtering techniques. IEEE Commun Surv Tutor 15(4):2070–2090
2. Anti Phishing Work Group (2014) Phishing attacks trends report. http://docs.apwg.org/reports/apwg_trends_report_q2_2014.pdf
3. Kumaraguru P, Rhee Y, Acquisti A, Cranor LF, Hong J, Nunge E (2007) Protecting people from phishing: the design and evaluation of an embedded training email system. In: CHI 2007: proceedings of the SIGCHI conference on human factors in computing systems, ACM, New York, pp 905–914
4. Sheng S, Magnien B, Kumaraguru P, Acquisti A, Cranor LF, Hong J, Nunge E (2007) Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In: SOUPS 2007: proceedings of the 3rd symposium on usable privacy and security, ACM, New York, pp 88–99
5. Sheng S, Wardman B, Warner G, Cranor LF, Hong J, Zhang C (2009) An empirical analysis of phishing blacklists. In: CEAS 2009
6. Almomani A, Gupta BB (2013) Phishing dynamic evolving neural fuzzy framework for online detection zero-day phishing E-mail. IJST 6(1):122–126
7. Zhang Y, Hong JI, Cranor LF (2007) Cantina: a content-based approach to detecting phishing web sites. In: Proceedings on WWW, ACM, New York, pp 639–648
8. Chen K-T, Huang C-R, Chen C-S (2010) Fighting phishing with discriminative key point features. IEEE Internet Community
9. Phishing URLs Dataset available at: https://www.phishtank.com

Author Biographies

Ankit kumar Jain is presently working as Assistant Professor in National Institute of Technology, Kurukshetra, India. He received Master of technology from Indian Institute of Information Technology Allahabad (IIIT) India. Currently, he is pursuing PhD in cyber security from National Institute of Technology, Kurukshetra. His general research interest is in the area of Information and Cyber security, Phishing Website Detection, Web security, Mobile Security, Online Social Network and Machine Learning. He has published many papers in reputed journals and conferences.

B. B. Gupta received Ph.D. degree from Indian Institute of Technology Roorkee, India in the area of Information and Cyber Security. He published more than 100 research papers (including 02 books and 14 book chapters) in International Journals and Conferences of high repute including IEEE, Elsevier, ACM, Springer, Wiley, Taylor & Francis, Inderscience, etc. His biography was selected and published in the 30th Edition of Marquis Who's Who in the World, 2012. Dr. Gupta also received Young Faculty research fellowship award from Ministry of Electronics and Information Technology, government of India in 2017. He is serving as associate editor of IEEE Access and Executive editor of IJITCA, Inderscience, respectively. He is also serving as guest editor of various reputed Journals. He was also visiting researcher with Yamaguchi University, Japan in January 2015. At present, Dr. Gupta is working as Assistant Professor in the Department of Computer Engineering, National Institute of Technology Kurukshetra India.