Journal of Southeast University (English Edition) Vol. 29, No. 2, pp. 134 - 138 June 2013 ISSN 1003—7985

Phishing detection method based on URL features

Cao Jiuxin[1,2] Dong Dan[1,2] Mao Bo[3] Wang Tianfeng[1,2]

([1]School of Computer Science and Engineering, Southeast University, Nanjing 211189, China)
([2]Key Laboratory of Computer Network and Information Integration of Ministry of Education, Southeast University, Nanjing 211189, China)
([3]Jiangsu Provincial Key Laboratory of E-Business, Nanjing University of Finance and Economics, Nanjing 210003, China)

Abstract: In order to effectively detect malicious phishing behaviors, a phishing detection method based on the uniform resource locator (URL) features is proposed. First, the method compares the phishing URLs with legal ones to extract the features of phishing URLs. Then a machine learning algorithm is applied to obtain the URL classification model from the sample data set training. In order to adapt to the change of a phishing URL, the classification model should be constantly updated according to the new samples. So, an incremental learning algorithm based on the feedback of the original sample data set is designed. The experiments verify that the combination of the URL features extracted in this paper and the support vector machine (SVM) classification algorithm can achieve a high phishing detection accuracy, and the incremental learning algorithm is also effective.

Key words: uniform resource locator (URL) features; phishing detection; support vector machine; incremental learning

doi: 10.3969/j.issn.1003-7985.2013.02.005

Phishing is the act of attempting to acquire information such as usernames, passwords, and credit card details (and sometimes, indirectly, money) by masquerading as a trustworthy entity in an electronic communication[1]. According to the phishing activity trends report of the 2nd quarter 2012 from the anti-phishing working group (APWG)[2], the total number of URLs used to host phishing attacks increased to 175 229, up from 164 023 in the first quarter of 2012. Financial services continued to be the most-targeted industry sector and followed by payment services. These phishing attacks have brought serious economic losses to the public. With the rapid development of Facebook, Twitter and other social networking sites, phishing behaviors have transited from stealing users' credit card account information to selling users' private information and other unlawful behaviors, which are becoming more serious with Trojan and Botnet technologies. Phishing has caused a great threat to Internet users, so how to detect phishing has become a hot research topic.

Domestic and foreign scholars have conducted a lot of research work on phishing detection. Chandrasekaran et al.[3] proposed a machine learning method based on mail structure characteristics. Cantina[4] was presented as a phishing detection tool through the analysis of the web content. Fu et al.[5] put forward a detection method on the basis of web image similarity. The method introduced the earth mover's distance (EMD) to calculate the visual similarity between the images, which performed well, but it only can be applied when the phishing page and the legitimate one are in a very similar image appearance. Cao et al.[6] also proposed an image-based detection algorithm. The algorithm calculated the similarity of web pages based on the attributed relational graph (ARG) of each page. Although these studies have made some progress in phishing detection, there are still some shortcomings such as weak universality and low efficiency in practical applications.

In consideration of these shortcomings, this paper proposes a URL feature-based phishing detection method. The method first compares the phishing URLs with legal ones to extract the features of phishing URLs. Then machine learning algorithms are used to obtain the URL classification model by training the sample data set, and the model is applied to detect the given URLs. In order to adapt to the change of a phishing URL, the classification model should be constantly updated according to the new samples, so an incremental learning algorithm based on the feedback of the original sample data set is designed.

1 Modeling Phishing URL Features

1.1 Phishing URL features analysis

In phishing attacks, the evildoers always try to absorb

Received 2012-12-10.

Biography: Cao Jiuxin (1967—), male, doctor, professor, jx.cao@seu.edu.cn.

Foundation items: The National Basic Research Program of China (973 Program) (No. 2010CB328104, 2009CB320501), the National Natural Science Foundation of China (No. 61272531, 61070158, 61003257, 61060161, 61003311, 41201486), the National Key Technology R&D Program during the 11th Five-Year Plan Period (No. 2010BAI88B03), Specialized Research Fund for the Doctoral Program of Higher Education (No. 20110092130002), the National Science and Technology Major Project (No. 2009ZX03004-004-04), the Foundation of the Key Laboratory of Network and Information Security of Jiangsu Province (No. BM2003201), the Key Laboratory of Computer Network and Information Integration of the Ministry of Education of China (No. 93K-9).

Citation: Cao Jiuxin, Dong Dan, Mao Bo, et al. Phishing detection method based on URL features [J]. Journal of Southeast University (English Edition), 2013, 29(2): 134 - 138. [doi: 10.3969/j.issn.1003-7985.2013.02.005]

victims into clicking a URL pointing to the phishing site. They usually obfuscate phishing URLs through various methods[7]. Every method attaches some features to the phishing URLs and these features can differentiate it from a legal one. Therefore, the URL features are essential to detect the phishing activities. By analyzing the phishing URLs we collect, the prominent features of a phishing URL are listed as follows:

1) Mixing IP address in the phishing URL. According to 3 000 phishing URLs and 1 000 legitimate ones, we can find that legitimate URLs containing an IP address almost do not exist.

2) Obfuscating the domain with a mass of dots. Phishing URLs usually use lots of dots to confuse users, for example, http://paypal.com.online-update.onlinebanking.service.customer./.... This kind of URL rarely exists in a legitimate URL.

3) Confusing users with abnormal depth of a URL path. In other words, there are many "/" in phishing URLs.

4) Confusing users with other special characters, such as "@", "~", "-". These special characters are often found in phishing URLs.

5) Abnormal numbers of the mixture of letters and digits in phishing URLs. This feature also appears in legitimate URLs, but it is more apparent in phishing ones.

6) Abnormal length of domain in the phishing URL. Under normal circumstances, the string appearing between the "http://" and the first "/" is considered as domain and the length is relatively longer in parts of phishing URLs.

7) Low PageRank. PageRank is a ranking of the pages recorded by Google according to their importance. We find that almost all the phishing URLs get low ranking or no record.

8) Suspicious words existing in the phishing URL. Some words appear frequently in phishing URLs, such as "login", "account", and the appearance locations of these items in phishing URLs has some difference with those in legitimate ones.

9) Imitating legitimate domain. For instance, the letter "1" in word "paypal" can be replaced with the digit "1", and the high similarity always successfully cheats users.

Of the above features, features 1) to 6) are easily obtained through the regular expression matching; feature 7) can be acquired by a third party (Google API); for feature 8), the frequency suspicious words can be achieved by applying the generalized suffix tree (GST); and the last feature 9) is more complex to obtain compared with the first eight, so further analysis will be focused on the imitating domain feature.

1.2 Calculating domain similarity

After analyzing the nearly 3 000 phishing URLs we have collected, it is found that only a few well-know domains (such as PayPal, Tibia, etc.) become the targets of more than half of URLs which have the feature of imitating domain. Most of these URLs just make a partial modification to confuse users. So, imitating legitimate domains is an important feature to detect the phishing URLs. One method to determine whether a domain imitates a legitimate one is to calculate the similarity of the two domains. The number of legitimate domains is so large that it is impractical to calculate the similarity with all legitimate ones. However, because of the concentration of the targeted legitimate domains, we can only calculate the similarity with the most targeted ones respectively and select the most similar ones.

In the field of biology, the famous algorithm to solve the gene sequence comparison problem is proposed by Smith and Waterman[8]. They applied dynamic programming to calculate the similarity of two gene sequences according to a pre-defined strategy, resulting in a similarity matrix $\boldsymbol{H}$. According to their method, the domain similarity matrix $\boldsymbol{H}$ in this paper is calculated as follows:

Suppose that the domain string extracted from the detected URL is $U = u_1 u_2 \cdots u_m$, the targeted domain string is $T = t_1 t_2 \cdots t_n$, the similarity matrix is defined as

$$H(i,0) = 0 \qquad 0 \leqslant i \leqslant m \tag{1}$$

$$H(0,j) = 0 \qquad 0 \leqslant j \leqslant n \tag{2}$$

$$\begin{aligned} H(i,j) = \max\{&0, H(i-1,j-1) + w(u_i,t_j), \\ &H(i-1,j) + w(\text{Deletion}), \\ &H(i,j-1) + w(\text{Insertion})\} \\ &1 \leqslant i \leqslant m,\ 1 \leqslant j \leqslant n \end{aligned} \tag{3}$$

where w is a series of pre-defined weight functions. There are four kinds of weight functions: matching function $w(\text{Match})$, non-matching function $w(\text{Mismatch})$, forward missing penalty function $w(\text{Deletion})$, reverse missing penalty function $w(\text{Insertion})$. In Eq. (3), if $u_i = t_j$ then $w(u_i,t_j) = w(\text{Match})$; otherwise, $w(u_i,t_j) = w(\text{Mismatch})$. $H(i,j)$, an element of matrix $\boldsymbol{H}$, indicates the similarity of the string $u_1 u_2 \cdots u_i$ and $t_1 t_2 \cdots t_j$. $H(m,n)$, the element in the bottom right corner of matrix $\boldsymbol{H}$, is the similarity of the string U and T. As the similarity is related with the length of the targeted domain string and pre-defined matching function, the normalizing process is needed to unify the similarity. $H'(m,n)$ can be used directly for similarity eigenvalue.

$$H'(m,n) = \frac{H(m,n)}{nw(\text{Match})} \tag{4}$$

2 Incremental Learning Algorithm Based on Support Vector Machine

In this paper, the phishing detection method based on URL features can separate the phishing ones from the le-

gitimate ones. However, the classification model is only trained one time on a large number of original sample sets, which does not have the ability of the evolvement with the ever-increasing sample sets. It means that the method lacks of the ability of incremental learning. Furthermore, the experiment in the third part shows that the support vector machine classification algorithm[9] is the most effective method in phishing detection. So, we present an incremental learning algorithm based on the support vector machine.

2.1 Problem description

The SVM-based incremental learning algorithm is described as follows:

- Prerequisite: Original sample set A and incremental sample set B, and $A \cap B = \varnothing$. ϕ^1 is the initial SVM classifier.
- Objective: To find the new SVM classifier trained from the sample set $A \cup B$.

2.2 Problem solution

The classical SVM algorithm does not support incremental learning. The simplest way to achieve the evolution of the classifier is called repeated learning (TISVM, for short), in which the new sample set is added into the original one and then repeat the learning process. This method is of low efficiency when the sample set reaches a certain size. The main idea of the traditional incremental learning algorithm based on SVM (SISVM, for short) is to retain the support vectors after obtaining a classifier, and to combine the existing vectors with the incremental samples as a new sample set[10-11]. Since the number of support vectors is much smaller compared with the original sample set, the training time is significantly reduced. In the aspect of accuracy, the classifier is comparable to the one trained by repeating the learning method when the distribution of incremental samples is in accordance with the original sample set. Otherwise, errors may be brought into the classifier[12], as shown in Fig. 1 and Fig. 2. Fig. 1 gives the initial classification result and Fig. 2 presents the new classification result after adding black incremental samples to the original ones. It can be easily found that the hyperplane of classification offsets and some of the original samples turn into support vectors with the help of new samples. Therefore, the incremental learning algorithm that only retains support vectors will inevitably bring errors to classification.

In order to reduce the error from the above algorithm, this paper designs an incremental learning algorithm based on feedback of the original sample data set (FISVM, for short). The main idea is that after obtaining the new classifier, we use it to identify the samples which do not agree with the new one from original samples. These samples may become support vectors, and they should be

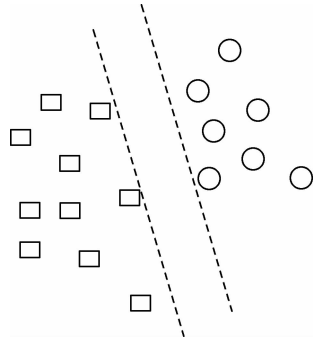

Fig. 1 Initial classification result

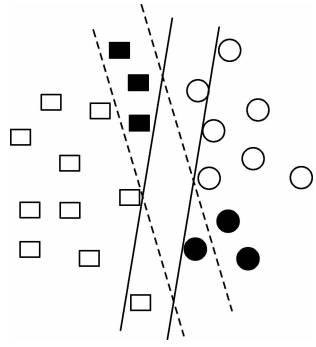

Fig. 2 New classification result after adding black incremental samples to the original ones

added into the new sample retraining process. The algorithm is listed as follows:

Algorithm 1 Incremental learning algorithm

```
/* Obtaining the classifier given the original sample set and the incremental sample set */
/* Sample is the defined type of sample data, and classifier is the defined type of SVM classifier */
classifier getClassifier(sample OriginalSet [ ], sample IncreSet [ ])
{
  /* C-SET is used to store training samples */
  sample C-SET[ ] = OriginalSet;
  /* B-SET is used to store the discard samples */
  sample B-SET[ ] = ∅;
  /* φH is the obtained classifier through training the C-SET; */
  classifier φH = SVMClassifier(C-SET);
  /* Verifying whether IncreSet accords with φH */
  identifyIncreSamples(IncreSet, φH, IncreSet^NK, IncreSet^K);
  if IncreSet^NK = ∅
    return φH;
  else
  {
    C-SET = C-SET ∪ IncreSet^NK;
    B-SET = B-SET ∪ IncreSet^K;
    classifier φ1 = SVMClassifier(C-SET);
    identifyIncreSamples(B-SET, φ1, IncreSet^NK, IncreSet^K);
```

```
        if IncreSet^NK = ∅
            return φ_1;
        else
        {
            C-SET = C-SET ∪ IncreSet^NK;
            classifier φ_e = SVMClassifier(C-SET);
            return φ_e;
        }
    }
}
```

3 Experimental Evaluation

In this section, we conduct experiments on the data set which is composed of phishing URLs collected in the well-known phishing identifying platforms (such as PhishTank) and the legitimate URLs collected in the Sougou corpus and Google's navigation website 265.com. First, we verify the feasibility of the phishing detection method, and then show the effectiveness of the proposed incremental learning algorithm based on SVM.

3.1 Phishing detection method evaluation

To minimize the overfitting which is common in machine learning, we make use of a 10 fold cross-validation method in the process of verification. Four kinds of machine learning algorithms are selected, which are the J48 classification tree, Naïve Bayes (NB), logistic regression (LR) and support vector machine (SVM), to carry out experiments on the same data set. The results are shown in Tab. 1, in which we can find that, in accuracy, SVM is the best, and, in a false positive rate, SVM is the lowest, and, in a false negative rate, SVM is in the second place following J48. With the comprehensive consideration from these three aspects, the SVM has the best performance. The performance of the four algorithms shows that the proposed phishing detection method is feasible and performs best combined with the SVM algorithm.

Tab. 1 Detection performance of four algorithms %

Algorithm	Accuracy	False positive	False negative
J48	93.3	5.4	1.8
NB	84.1	10.7	21.4
LR	93.4	6.7	6.6
SVM	95.5	4.6	4.4

3.2 Incremental learning algorithm evaluation

In order to verify the validity of the incremental learning algorithm, we compare the proposed incremental learning algorithm (FISVM) with the traditional incremental learning algorithm (TISVM) and the incremental learning algorithm based on the support vectors (SISVM).

In the experiment, the initial samples contain 300 phishing URLs and 300 legitimate ones, and 400 incremental samples are added to the initial samples at each incremental step with 500 samples for testing. The results are shown in Fig. 3, which indicates that the TISVM achieves the highest classification accuracy in each incremental step benefiting by making use of information of all the samples. The characteristics of the SISVM makes the training set small so that the complexity is low but the accuracy is ineffective and volatile. Although the precision of the FISVM is lower than that of the TISVM, the overall trend is consistent with the TISVM. Moreover, it has small fluctuations. In summary, our incremental learning algorithm FISVM is close to the TISVM in accuracy and has good stability.

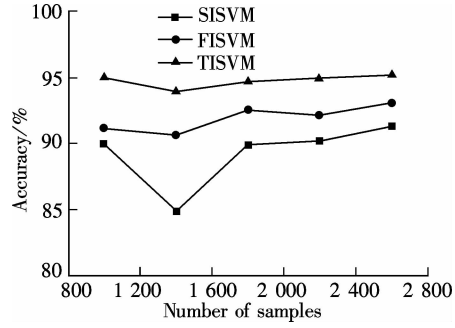

Fig. 3 Accuracy of the incremental learning algorithm

4 Conclusion

In this paper, we propose a phishing detection method using the URL as the entry point. Meanwhile, in order to adapt to the change of phishing URLs, the classification model should be constantly updated according to the new samples. So, an incremental learning algorithm based on the feedback of the original sample data set is designed. In experiments, we compare four machine learning algorithms to verify the phishing detection method and simultaneously verify the effectiveness of the incremental learning algorithm. The results show that the proposed phishing detection method is feasible and performs best combined with the SVM algorithm, and the incremental learning algorithm is also effective.

References

[1] Wikipedia. Phishing[EB/OL]. (2013-04-20)[2013-04-27]. http://en.wikipedia.org/wiki/Phishing.

[2] Anti-Phishing Working Group. Phishing activity trends report [EB/OL]. (2012-10-17)[2013-03-16]. http://docs.apwg.org/reports/apwg_trends_report_q2_2012.pdf.

[3] Chandrasekaran M, Narayanan K, Upadhyaya S. Phishing email detection based on structural properties[C]// *NYS Cyber Security Conference*. New York, USA, 2006: 2-8.

[4] Zhang Y, Hong J I, Cranor L F. Cantina: a content-based approach to detecting phishing web sites[C]//16*th International World Wide Web Conference*. Banff, Alberta, Canada, 2007:639-648.

[5] Fu A Y, Liu W Y, Deng X T. Detecting phishing web pages with visual similarity assessment based on earth mover's distance (EMD) [J]. *IEEE Transactions on Dependable and Secure Computing*, 2006, **3**(4): 301 - 311.

[6] Cao Jiuxin, Mao Bo, Luo Junzhou, et al. A phishing web pages detection algorithm based on nested structure of earth mover's distance (nested-EMD) [J]. *Chinese Journal of Computers*, 2009, **32**(5): 922 - 929. (in Chinese)

[7] Garera S, Provos N, Chew M, et al. A framework for detection and measurement of phishing attacks[C]//*Proceedings of the* 2007 *ACM Workshop on Recurring Malcode*. Alexandria, VA, USA, 2007: 1 - 8.

[8] Smith T F, Waterman M S. Identification of common molecular subsequences [J]. *Journal of Molecular Biology*, 1981, **147**(1): 195 - 197.

[9] Chang C C, Lin C J. LIBSVM: a library for support vector machines [J]. *ACM Transactions on Intelligent Systems and Technology*, 2011, **2**(3): 1 - 27.

[10] Domeniconi C, Gunopulos D. Incremental support vector machine construction[C]//*Proceedings of IEEE International Conference on Data Mining*. San Jose, CA, USA, 2001: 589 - 592.

[11] Syed N A, Liu H, Sung K K. Incremental learning with support vector machines[C]//*Proceedings of the Workshop on Support Vector Machines at the International Joint Conference on Artificial Intelligence*. Stockholm, Sweden, 1999: 876 - 892.

[12] Wang W J. A redundant incremental learning algorithm for SVM[C]//*Proceedings of the 7th International Conference on Machine Learning and Cybernetics*. Kunming, China, 2008: 734 - 738.

基于 URL 特征的 Phishing 检测方法

曹玖新[1,2] 董 丹[1,2] 毛 波[3] 王田峰[1,2]

([1] 东南大学计算机科学与工程学院,南京 211189)

([2] 东南大学网络和信息集成教育部重点实验室,南京 211189)

([3] 南京财经大学江苏省电子商务重点实验室,南京 210003)

摘要:为了有效检测恶意网络钓鱼(phishing)行为,提出一种基于 URL 特征的 phishing 检测方法. 该方法首先对现有钓鱼 URL 与合法 URL 进行分析对比,提取钓鱼 URL 的显著特征,然后采用机器学习算法对样本数据集训练从而获得分类检测模型,用来检测待检测的 URL. 为适应钓鱼 URL 的变化,分类模型需要根据新增样本不断更新,因此,设计了一种基于原始样本数据反馈的增量学习算法. 实验表明:提取的 URL 特征与支持向量机(SVM)分类算法的结合能够使 phishing 检测达到较高的检测精度,且该增量学习算法是有效的.

关键词:URL 特征;phishing 检测;支持向量机;增量学习

中图分类号:TP393