http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 11 (7): 921-925, 2012
ISSN 1812-5638 / DOI: 10.3923/itj.2012.921.925

A SVM-based Technique to Detect Phishing URLs

Huajun Huang, Liang Qian and Yaojun Wang
College of Computer and Information Engineering,
Central South University of Forestry and Technology, Changsha 410004, China

Abstract: Phishing, a term coined in 1996, is a form of online identity theft. Phisher tries to lure her victim into clicking a phishing URL pointing to a spoof page via spam-email to harvest financial information. The phishing activity is on the rise and their techniques become easier and more sophisticated. Quite a number of solutions to mitigate phishing attacks have been proposed to date. Those methods fetch webpage content which result in undesired side effects. In this paper, a novel method is proposed to detect phishing URL based on SVM. The feature vector is constructed with 23 features to model the SVM which 4 features are the structure feature of the phishing URL, 9 features are lexical feature and 10 features are mostly target phished brand name of website. The experimental results show the detection solution achieves 99.0% accuracy on average that the phishing URLs achieve is downloaded in PhishTank.

Key words: Phishing, phishing URLs, SVM, feature vector

INTRODUCTION

Phishing is a form of online identity theft (APWG, 2008). Phishers lure unsuspecting victims into counterfeited websites designed to trick recipients' confidential information with spoofed e-mails. The confidential information include user names and passwords, social security numbers, credit card numbers, bank account numbers and personal information such as birthdates and mothers' maiden names (Sudha *et al*., 2007). Even though most attacks are surprisingly straight-forward, for example phisher asking a victim for his bank account number and PIN, they are also rather successful.

Anti-phishing is the countermeasure solution to defeat phishing. There is quite a number of anti-phishing proposed countermeasures to date. Generally speaking, past works in anti-phishing can be classified into phishing detection (Kirda and Kruegel, 2006), phishing email filter (Al-Momani *et al*., 2011), tracking phishing site (Zhou *et al*., 2009), phisher behave analysis (McGrath and Gupta, 2008), take down phishing site host (Moore and Clayton, 2007) and so on. Detection is a very important aspect in the fight against phishing.

One of those is a browser-side technique. Browser-side-based solutions embed anti-phishing measures 'plug-in' into Web browsers. These browsers use web pages' visual behaviors to prevent cheating. According to the approaches used in browser side, we roughly divide them into four categories: phishing URL detection approaches (Ma *et al*., 2011; Thomas *et al*., 2011; Blum *et al*., 2010), heuristic-based detection method (Zhang *et al*., 2007), visual similarity anti-phishing solution (Fu *et al*., 2006).

In this study, we focus on the phishing URL detection method. This method exploits the anatomy of phishing URLs structure, lexical features used in URL, domain name always spoofed by phisher and phishing site host information, to indicate the suspicious URL belongs to a phishing site. Phishing URL detection solution doesn't require any knowledge of the corresponding webpage content. Existing anti-phishing methods, whether heuristic based or visual similarity based, fetch webpage content which result in undesired side effects, such as signing up to mailing list or even acknowledging receipt of a credit card. The phishing URL classification scheme based only on examining the suspicious URL can avoid unwanted events to the end user.

In this study, a novel method is proposed to detect phishing URL based on SVM. Firstly, we exploit this observation of heuristics in the structure of URL, the lexical feature in URL characters and the phishing target brand name. The feature vector is constructed with 23 features to model the SVM which 4 features are the structure feature of the phishing URL, 9 features are lexical feature and 10 features are brand name of website. Lastly, a lot of experiments are done. The experimental results the detection method achieves 99.0% accuracy on average, that the phishing URLs achieve is downloaded in PhishTank.

Corresponding Author: Huajun Huang, College of Computer and Information Engineering,
Central South University of Forestry and Technology, Changsha 410004, China

SVM THEORY

In this study, we use SVM theory in LIBSVM (Chang and Lin, 2011). Given training vectors $x_i \epsilon R^n$, i = 1,....,l, in two classes and a vector $y \epsilon R^l$ such that $y_i \epsilon \{1, -1\}$, the primal problem is defined as follows:

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \\ &\text{s.t.}\, y_i \left(w^T \phi(x_i) + b \right) \geq 1 - \xi_i \\ &\xi_i \geq 0, i = 1, \ldots, l \end{aligned} \tag{1}$$

Its dual is:

$$\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha - e^T \alpha \tag{2}$$

$$\text{s.t. } y^T, \alpha = 0$$

$$0 \leq \alpha_i \leq C, \ i = 1, \ldots, l,$$

where, e is the vector of all ones, C is the upper bound, where C>0; Symbol Q is an l×l positive semidefinite matrix and the functions $Q_{ij} = y_i y_j K(x_i, x_j)$ and $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ are the kernel. Using the function ϕ, the training vectors x_i are mapped into a higher dimensional space.

The decision function is written in the following:

$$\text{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i K(x_i, x) + b \right) \tag{3}$$

Using SVM to decide the phishing URL, when a suspicious URL is labeled asy_i = -1, the URL is phishing URL, or y_i = 1 is labeled, the suspicious URL is legal.

PROPOSED SOLUTION

Figure 1 shows the flow of detecting phishing URL. The system is consisted two stages: training stage and classifier stage. In the training stage, 23 feature values are extracted from instance in the training achieve. The feature vector are organized in LIBSVM proper format to find the optimal parameters used in LIBSVM. At the classifier stage, features values and feature vector format are the same to the training stage. The output label values indicate the input suspicious URL is phishing URL or not. When the label is equal to "-1", the suspicious URL is phishing URL and equal to "1", the suspicious URL is belong to non-phishing class. URL (uniform resource locator) is used to locate web sites and individual web resource. URL has the following standard syntax.

<protocol> ://< hostname><path>

The *<protocol>* portion indicates which network protocol will be used to fetch the requested resource. The *<hostname>* is the identifier of the Web server. The *<path>* of a URL is analogous to the path name of a file on a local computer.

A phisher usually lure the end-user into clicking an obfuscated URL pointing to the phishing site. Figure 2 shows four obfuscation techniques to make phishing URL. Some related works have examined the statistics of an obfuscated phishing URLs in some way. Our solution in this article is different to previous work in the following respects: we exploit 4 URL's structure features, 9 lexical features and 10 brand name features, we train the phishing URL and non-phishing URL feature vector in MATLAB using LIBSVM tool.

Structure features: In this portion, we use a combination of features described by McGrath and Garera. We select

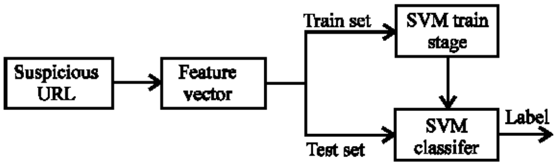
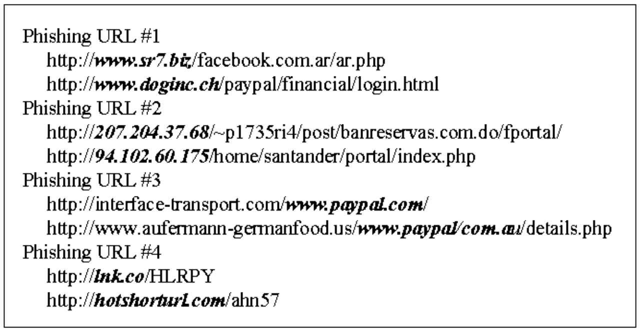


Fig. 1: The flow of detecting phishing URL



Phishing URL #1
http://***www.sr7.biz***/facebook.com.ar/ar.php
http://***www.doginc.ch***/paypal/financial/login.html
Phishing URL #2
http://***207.204.37.68***/~p1735ri4/post/banreservas.com.do/fportal/
http://***94.102.60.175***/home/santander/portal/index.php
Phishing URL #3
http://interface-transport.com/***www.paypal.com***/
http://www.aufermann-germanfood.us/***www.paypal/com.au***/details.php
Phishing URL #4
http://***lnk.co***/HLRPY
http://***hotshorturl.com***/ahn57

Fig. 2: The structure features of phishing URL

IP address, the length of hostname, the number of dots in the <*path*> part of URL, the number of dash in the <*hostname*> portion of URL. We use F1, F_2, F_3 and F_4 to indicate these features in sequence.

Lexical features: Phishing URLs tend to "look different" in the eyes of the end-users which is the justification for using lexical features to distinguish URLs to malicious sites. As Garera proposed in previous studies, they selected the tokens, such as confirm, banking, secure, ebayisapi, webscr, log in, sign in, as lexical features. We also find the token of "http" often appears in the <path> part of labeled phishing URLs in PhishTank So, we take the word token, http, confirm, banking, secure, ebayisapi, webscr, log in, sign in, as lexical features and denotes as F_5, F_6, F_7, F_8, F_9, F_{10}, F_{11}, F_{12} and F_{13}.

Brand name features: We observed that phishing URLs in order to lure the victims to will contain several suggestive word tokens which is always the targeted site's brand name. But this phenomenon is difficult happen in benign URLs. In PhishTank data achieve, we analysis is several monthly stats archive and select top 10 brand names listed in PhishTank in July 2011 stats archive. The 10 brand names are eBay, PayPal, sulake, facebook, orkut, santander, mastercard, warcraft, visa, bradesco and the symbols F_{14}, F_{15}, F_{16}, F_{17}, F_{18}, F_{19}, F_{20}, F_{21}, F_{22} and F_{23} are denoted brand name features.

At last, we get the feature vector as follows:

$$x_i = <F_1, F_2, F_3, F_4, F_5, F_6, F_7, F_8, F_9, F_{10}, F_{11}, F_{12}, F_{13}, F_{14}, F_{15}, F_{16}, F_{17}, F_{18}, F_{19}, F_{20}, F_{21}, F_{22}, F_{23}> \quad (4)$$

In our solution, the feature value equal to 1 mean that it is phishing feature, 0 is non-phishing feature. Next we show how to set value to feature vector.

The value of structure features, F_1, F_2, F_3 and F_4 is set with formula 5, 6, 7 and 8:

$$F_1 = \begin{cases} 1, \text{ if URL is IP address} \\ 0, \text{ others} \end{cases} \quad (5)$$

$$F_2 = \begin{cases} 1, \text{ if length (hostname)>22} \\ 0, \text{ others} \end{cases} \quad (6)$$

$$F_3 = \begin{cases} 1, \text{ if dot (path)>2} \\ 0, \text{ others} \end{cases} \quad (7)$$

$$F_4 = \begin{cases} 1, \text{ if dash (hostname)>2} \\ 0, \text{ others} \end{cases} \quad (8)$$

The function length (hostname) in formula 6 calculates the length of the <hostname> of URL. The function dot (Path) counts the number of the dot, ".", in <path> of URL. Function dash (hostname) returns the number of dash, "-", in the <hostname> of URL.

The value of lexical feature and brand name feature are used the same formula 9:

$$F_i = \begin{cases} 1, \text{ URL contain w} \\ 0, \text{ others} \end{cases} \quad (9)$$

where, $4 \le i \le 23$ and w = {http, confirm, banking, secure, ebayisapi, webscr, log in, sign in, eBay, PayPal, sulake, facebook, orkut, santander, mastercard, warcraft, visa, bradesco}.

EXPERIMENTAL RESULTS AND ANALYSIS

The labeled phishing URLs is downloaded from PhishTank. At Sept. 2 2011, we downloaded 5218 online phishing URLs from, denoted as DS1. At Sept. 18 2011, another labeled phishing URLs data set, denoted as DS2, was got from Phishing which contained 4876 phishing URLs. For the non-phishing achieve, we chose from the open directory, such as Yahoo and DMOZ directory. At Sept. 5 2011, we collected 2099 non-phishing URLs from two sites, denoted as DS3.

The feature extraction algorithm is implemented with Java and the classification solution is designed in MATLAB with LIBSVM tool. Feature vectors are stored as the rows of a sparse matrix.

Firstly, we analysis the domain's length of phishing and non-phishing URL in data set DS1 and DS3. Figure 3 shows the distribution of domain name in DS1 and the average length is 22 characters. The distribution of domain name length is plotted in Fig. 4 and the average length is 15 characters. We also found that the length of domain name in non-phishing data set almost less than 22 characters.

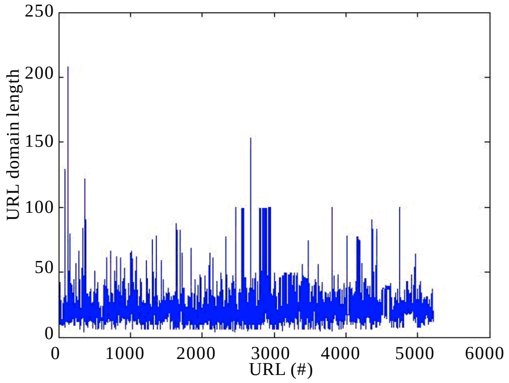


Fig. 3: Phishing URL domain length (DS1)

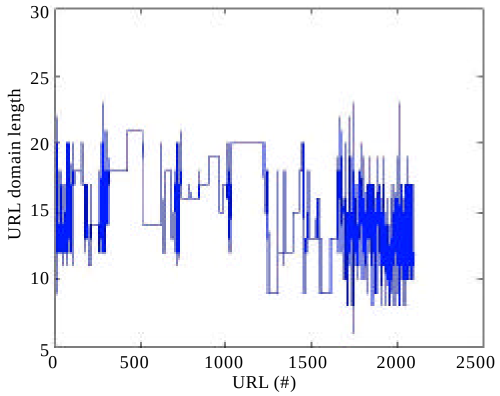


Fig. 4: Non-phishing URL domain length (DS3)

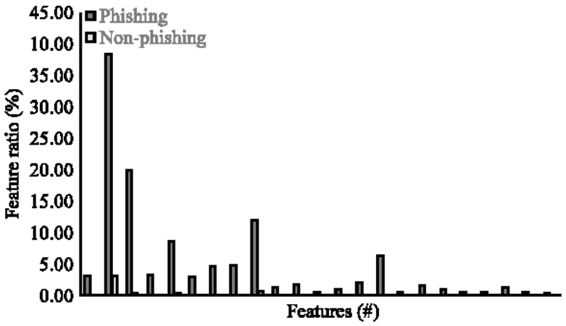


Fig. 5: Phishing and non-phishing feature accuracy ratio (DS2, DS3)

Next, we verify each feature accuracy ratio in data set DS2 and DS3. We also contrast the feature ratio of DS2 to DS3 and plot in Fig. 5. In this Fig. 5, we find that each chosen feature happened in DS2 and the most accuracy ratio feature is $F_{2,}$ obtained to 38.74%. But to DS2, most of features' accuracy ratio is zero, except for feature F_2, F_3, F_5 and F_9.

In training stage, we randomly selected 2963 feature vectors from data set DS2 and 1143 feature vectors from DS to construct the train set. After this stage, the classifier can correct classify 4069 feature vector in 4106 and the detection accuracy is achieved 99.1%. To show the train stage effectiveness, the ROC curve (receiver operator characteristic curve, ROC) is shown in Fig. 6. The area under the ROC curve for the positive class is 0.99565.

At last, we verify the effectiveness of the solution. 3137 feature vectors are randomly selected in phishing achieve DS1 and 1866 feature vectors are randomly selected in data set DS3. The false negative (FN) is that a phishing URL is classified into non-phishing URL class. The false positive (FP) is that a non-phishing URL is classified into phishing URL class. To data set DS1, we test the false negative and the false positive is done in data set DS3. Figure 7 shows he ROC curve of DS1 and DS3. The area under the ROC curve for the positive class is 0.99705, thus confirming that our classifier has a high accuracy of phishing URL detection.

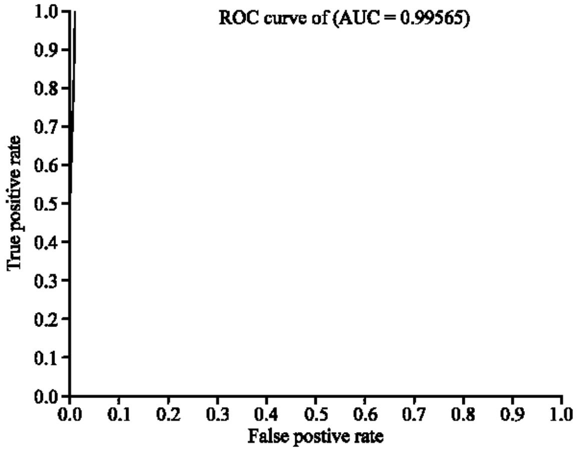


Fig. 6: ROC of train set (DS2+DS3)

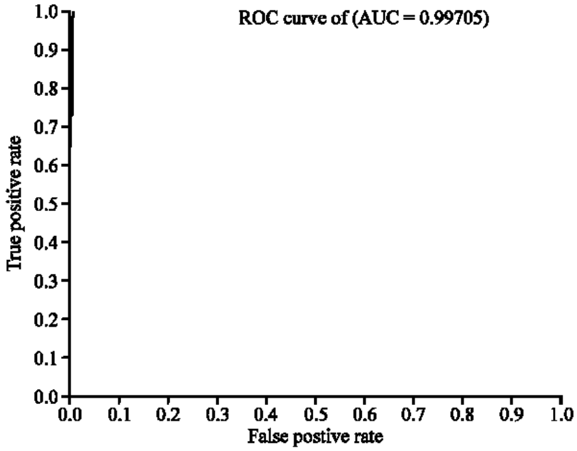


Fig. 7: ROC of test set (DS1+DS3)

Some previous works are proposed to detect phishing URLs. The method proposed by Ma *et al.* (2011) is similar to our. As depicted, they got 99% accuracy with 150,000 features. Contrast to Ma's, we also archive average 99% accuracy with only 23 features. So the execute effectiveness is better than Ma's. Another method is of Blum *et al.* (2010) which is only use large lexical features to train model. As they say, the cumulative error rate is as low as 3%. Using FN add FP, the cumulative error rate is only 2.0%, less than Blum's.

CONCLUSION

Phishing is an important problem that results in identity theft. Although simple, phishing attacks are highly effective and have caused billions of dollars of damage in the last couple of years. In many cases, the phisher does not directly cause the economic damage but resells the illicitly obtained information on a secondary market. Hence, phishing attacks are still important and solutions of the problems are required.

In this research, we study the structure of URL, the lexical feature in URL characters and the phishing target brand name and propose a SVM-based phishing URL detection solution. The experimental results show the solution is effective to catch phishing URLs and used as plug-in in browser to filter the phishing site.

ACKNOWLEDGMENTS

This study is supported by Hunan Provincial Natural Science Foundation of China (No. 10JJ4043, 10JJ5062); Scientific Research Fund of Hunan Provincial Education Department (No. 08B091); Hunan Provincial Science and Technology Major Project (No. 2010J05); Hunan Province Planned Science and Technology Key Project (No. 2010NK2003) and Hunan Province Planned Science and Technology Project (No. 2010TZ4012).

REFERENCES

APWG, 2008. Anti-phishing working group. http://www.antiphishing.org

Al-Momani, A.A.D., T.C. Wan, K. Al-Saedi and A. Altaher, 2011. An online model on evolving phishing e-mail detection and classification method. J. Applied Sci., 11: 3301-3307.

Blum, A., B. Wardman and T. Solorio, 2010. Lexical feature based phishing URL detection using online learning. Proceedings of the 3rd Workshop on Artificial Intelligence and Security, October 2010, Chicago, Illinois, USA., pp: 54-60.

Chang, C.C. and C.J. Lin, 2011. LIBSVM: A library for support vector machines. ACM Trans. Intellig. Syst. Technol., 3: 1-27.

Fu, A.Y., L. Wenyin and X. Deng, 2006. Detecting phishing web pages with visual similarity assessment based on earth mover`s distance (EMD). IEEE Trans. Dependable Secure Comp., 3: 301-311.

Kirda, E. and C. Kruegel, 2006. Protecting users against phishing attacks. Comp. J., 49: 554-561.

Ma, J., L.K. Saul, S. Savage and G.M. Voelker, 2011. Learning to detect malicious URLs. ACM Trans. Intellig. Syst. Technol., 2: 1-23.

McGrath, D.K. and M. Gupta, 2008. Behind phishing: An examination of phisher modi operandi. Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats, April 15, 2008, San Francisco, CA, USA., pp: 1-8.

Moore, T. and R. Clayton, 2007. The impact of incentives on notice and take-down. Proceedings of the 7th Workshop on the Economics of Information Security, October 4-5, 2007, Pittsburgh, PA., USA., pp: 1-24.

Sudha, R., A.S. Thiagarajan and A. Seetharaman, 2007. The security concern on internet banking adoption among Malaysian banking customers. Pak. J. Biol. Sci., 10: 102-106.

Thomas, K., C. Grier and J. Ma, 2011. Design and evaluation of a real-time URL spam filtering service. Proceedings of the IEEE Symposium on Security and Privacy, May 22-25, 2011, Berleley, California, USA., pp: 16-31.

Zhang Y., J. Hong and L. Cranor, 2007. CANTINA: A content-based approach to detecting phishing web sites. Proceedings of the 16th International World Wide Web Conference, May 8-12, 2007, Banff, Alberta, Canada, pp: 639-648.

Zhou, C.V., C. Leckie and S. Karunasekera, 2009. Collaborative detection of fast flux phishing domains. J. Networks, 4: 75-84.