Malicious Web Pages Detection Using Static Analysis of URLs

Dharmaraj R. PATIL, J.B. PATIL
Department of Computer Engineering, RCPIT, Shirpur, India
dharmaraj.rcpit@gmail.com, jbpatil@hotmail.com

Abstract

Malicious Web pages detection becomes a crucial task, due to the ever changing nature of attacks and the structure of todays Web pages. Attackers employ different techniques for attack constructions. Therefore, feature selection and dataset preparation plays an important role in detection of malicious Web pages. While the existing approaches provide a promising solution in detection of malicious Web pages, still there are open issues in the effective detection. In this paper, we have provided the static analysis of URL string for effective detection of malicious Web pages. We have considered only static features of Web page URLs. We have extracted 79 static features of URLs and domain names from benchmarks benign and malicious URLs. We have evaluated several batch learning algorithms like SVM, AdaBoost, J48, Random Forest, Random Tree, Naive Bayes, Logistic Regression, SGD and BayesNet on our dataset. Our experimental analysis shows promising detection results, a detection rate between 95%-99% and very low false positive rate (FPR) and false negative rate (FNR) for all the classification models.

Index terms: Feature extraction, Machine learning, Static analysis, Security, URLs

References:

[1]. N. Provos, P. Mavrommatis, M. A. Rajab and F. Monrose, "All Your iFRAMEs Point to Us", In Proc. of the 17th conference on Security Symposium, SS'08,USENIX Association Berkeley, CA, USA, 2008, pp. 1-15.

[2]. B. Liang, J. Huang, F. Liu, D. Wang, D. Dong and Z. Liang, "Malicious Web Pages Detection Based on Abnormal Visibility Recognition", International Conference on E-Business and Information System Security, EBISS '09, Wuhan, 2009, pp. 1-5.

[3]. D. Canali, M. Cova, C. Kruegel and G. Vigna, "Prophiler: A fast filter for the large-scale detection of malicious Web pages", In Proc. of the 20th International Conference on World Wide Web (WWW 11), Hyderabad, India, 2011, pp. 197-206.

[4]. Dharmaraj Rajaram Patil and J. B. Patil, "Survey on Malicious Web Pages Detection Techniques", International Journal of u- and e- Service, Science and Technology (IJUNESST), Vol.8, No.5, 2015, pp.195-206.

[5]. J. Ma, L. Saul, S. Savage and G. Voelker, "Learning to Detect Malicious URLs", ACM Transactions on Intelligent Systems and Technology, New York, USA ,Vol. 2, No. 3, Article 30, 2011, pp. 30:1-30:24.

[6]. Hyunsang Choi, Bin B. Zhu and Heejo Lee, "Detecting Malicious Web Links and Identifying Their Attack Types", In Proc. of the 2nd USENIX conference on Web application development (WebApps'11), USENIX Association Berkeley, CA, USA, 2011, pp. 1-12.

[7]. Christian Seifert, Ian Welch and Peter Komisarczuk," Identification of Malicious Web Pages with Static Heuristics", In Proc. of the Telecommunication Networks and Applications Conference (ATNAC 2008), Adelaide, SA, 2008, pp. 91 - 96.

[8]. Birhanu Eshete, Adolfo Villafiorita and Komminist Weldemariam, "BINSPECT: Holistic Analysis and Detection of Malicious Web Pages", In Proc. of the 8th International ICST Conference, SecureComm 2012, Padua, Italy, 2012, pp. 149-166.

[9]. P. Likarish, E. Jung and I. Jo, "Obfuscated Malicious JavaScript Detection using Classification Techniques", In Proc. of the 4th International Conference on Malicious and Unwanted Software (MALWARE), Montreal, QC, 2009, pp. 47-54.

[10]. W. Zhang, Y. Ding, Y. Tang and B. Zhao, "Malicious Web page detection based on on-line learning algorithm", In Proc. of the International Conference on Machine Learning and Cybernetics (ICMLC), Guilin, Vol.4, 2011, pp. 1914-1919.

[11]. A. Le, A. Markopoulou and M. Faloutsos, "PhishDef: URL Names Say It All", In Proc. of IEEE INFOCOM, Shanghai, 2011, pp. 191-195.

[12]. Malc0de Blacklist. Available: http://malc0de.com/bl/; accessed on 20 December 2014.

[13]. MalwareDomainList.com Hosts List. Available: http://www.malwaredomainlist.com/hostslist/hosts.txt/; accessed on 20 December 2014.

[14]. Sujata Garera, Niels Provos, Monica Chew and Aviel D. Rubin, "A Framework for Detection and Measurement of Phishing Attacks", In Proc. of the ACM Workshop on Recurring Malcode (WORM '07), Alexandria, Virginia, USA, 2007, pp. 1-8.

[15]. StopBadware Top 50 IP. Available: https://www.stopbadware.org/top-50#top_ip/; accessed on 20 December 2014.

[16]. DNS-BH – Malware Domain Blocklist. Available: https://www.malware-domains.com/files/domains.zip/; accessed on 25 December 2014.

[17]. Weka 3: Data Mining Software in Java. Available: http://www.cs.waikato.ac.nz/ml/weka/; accessed on 10 December 2015.

[18]. Saeed Abu-Nimeh , Dario Nappa , Xinlei Wang and Suku Nair, "A Comparison of Machine Learning Techniques for Phishing Detection", In Proc. of the anti-phishing working groups 2nd annual eCrime researchers summit, eCrime '07, Pittsburgh, PA, USA, 2007, pp. 60-69.

[19]. LIBLINEAR – A Library for Large Linear Classification. Available: http://www.csie.ntu.edu.tw/ cjlin/liblinear/; accessed on 10 January 2013.

[20]. Robert E. Schapire, "Explaining AdaBoost", Princeton University, Dept. of Computer Science, 35 Olden Street, Princeton, NJ 08540 USA, pp. 1-16.

[21]. J48 Decision Trees. Available: http://www.d.umn.edu/~padhy005/Chapter5.html/; accessed on 18 February 2016.

[22]. Naive Bayesian. Available: http://www.saedsayad.com/naive_bayesian.htm/; accessed on 18 February 2016.

[23]. J. Ma, L. Saul, S. Savage and G. Voelker, "Beyond Blacklists: Learning to Detect Malicious Websites from Suspicious URLs", In Proc.of the 15th ACM SIGKDD International Conference on Knowledge discovery and data mining, KDD '09, New York, NY, USA, 2009, pp. 1245-1254.

[24]. PhishTank: Phishtank developer information. Available: http://www.phishtank.com/developer info.php/; accessed on 25 November 2014.

[25]. OpenPhish - Phishing Intelligence Feeds. Available: https://openphish.com/; accessed on 25 November 2014.

[26]. MalwareURL: Malware urls. Available: http://www.malwareurl.com/; accessed on 25 November 2014.

[27]. Alexa: Alexa top global websites. Available: http://www.alexa.com/topsites/; accessed on 10 December 2014.

[28]. DMOZ: Open directory project. Available: http://www.dmoz.org/; accessed on 20 December 2014.

[29]. Rong-En Fan, Kai-Wei Chang, ho-Jui Hsieh, Xiang-Rui Wang and Chih-Jen Lin, "LIBLINEAR: A Library for Large Linear Classification", Journal of Machine Learning Research, Vol. 9, 2008, pp. 1871-1874.