International Journal of Security and Its Applications
Vol.8, No.1 (2014), pp.291-294
http://dx.doi.org/10.14257/ijsia.2014.8.1.27

Risk Prediction of Malicious Code-Infected Websites by Mining Vulnerability Features

Taek Lee[1], Dohoon Kim[2], Hyunchoel Jeong[3] and Hoh Peter In[1,*]

[1]*College of Information and Communications, Korea University*
[2]*Agency for Defense Development*
[3]*Korea Internet and Security Agency*
comtaek@korea.ac.kr, karmy01@gmail.com, hcjung@kisa.or.kr, hoh_in@korea.ac.kr

Abstract

Malicious-code scanning tools are practically available for identifying suspicious websites. However, such tools only warn users about suspicious sites and do not provide clues as to why the sites were hacked and which vulnerability was responsible for the attack. In addition, the huge number of alarms burdens mangers while executing in-time-response duties. In this paper, a process involving feature modeling and data-mining techniques is proposed to help solve such problems.

Keywords: *feature modeling, classification, vulnerability identification*

1. Introduction

Malicious-code scanning systems [1] are widely used to prevent computers of innocent users from being infected by malicious codes via web pages tainted by hackers. However, the problem with the countermeasures is that they are only reactive responses. The huge number of alarm logs makes the situation worse because these calls cannot be treated individually and appropriately within the required time.

To develop proactive responses, it is important to ascertain the vulnerabilities that are frequently exploited and prioritize those to be dealt with first in triage. In this paper, we propose feature modeling of the vulnerabilities of websites infected by malicious codes in order to identify significant vulnerabilities and understand their quantitative impact. Even though feature modeling study [2] had been tried to classify suspicious website URLs, the study does not address vulnerability impact.

2. The Proposed Process for Identifying Vulnerability Features

As shown in Figure 1, this section presents detailed steps for quantifying the severity of security risks posed by websites, labeling binary classes (high/low) on the basis of the quantified scores, and testing the impact of vulnerabilities on the severity classification.

* Hoh Peter In is the corresponding author

ISSN: 1738-9976 IJSIA

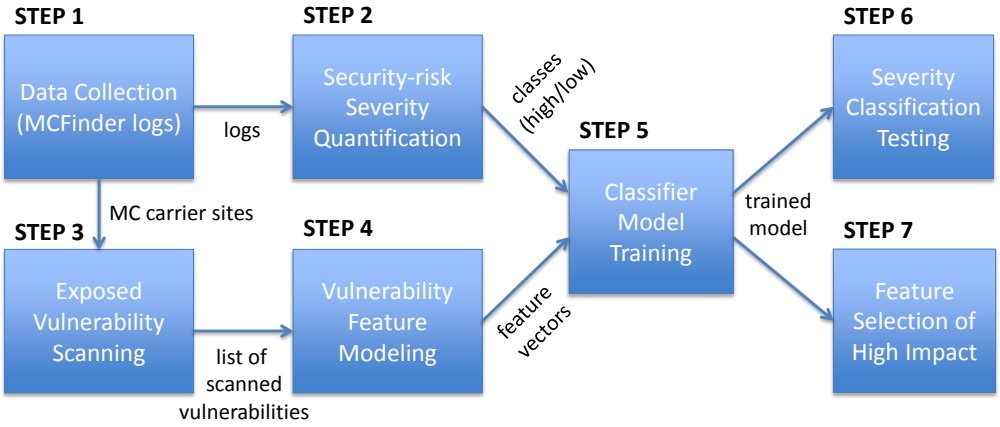


Figure 1. Overview of the Proposed Process

2.1. Data Collection by Malicious Code Finder (MCFinder) – STEP 1

MCFinder [1] is a malicious-code detection system developed and operated by krCERT[1]. It detects embedded malicious codes by scanning input lists of suspicious websites. Consequently, it finds tainted websites used for trapping purposes (*i.e.*, malicious-code carrier sites) and seed sites that spread malicious codes (*i.e.*, malicious-code distribution sites). Basically, these logs of the detected sites are used as a source of information for data mining.

2.2. Severity Quantification and Class Labeling – STEP 2

By studying the entries in the MCFinder logs, the severity of the security risk of malicious-code carrier sites can be measured. To score the severity level of the malicious-code carrier sites, the product of the following two factors are taken into account: the number of detections recorded in history and the number of sibling sites tainted, i.e., the sites commonly linked to a certain malicious-code distribution site. The greater the number of detection records and similarly exploited cases, higher is the score and the more severe the risk is considered. After sorting the scores, the top 25% and bottom 25% sites are respectively labeled as high severity class and low severity class sites.

2.3. Survey of Exposed Vulnerabilities – STEP 3

To survey potential vulnerabilities of victims sites (i.e., malicious-code carrier sites), the vulnerability searching tool Nikto[2] is used to determine whether an input site has any certain vulnerability listed in the Open Source Vulnerability Database (OSVDB)[3].

[1] http://www.krcert.or.kr
[2] http://cirt.net/nikto2
[3] http://www.osvdb.org

2.4. Vulnerability Feature Modeling – STEP 4

Our feature modeling is based on the Bag-of-Words (BOW) [3] model that is popularly used in document classification. In this model, a document is represented as a collection of words. However, in our context, a document means a website with inherent vulnerabilities and words mean vulnerabilities. Thus, once any vulnerability has been searched for, the number of counted searches is recorded as a word-count element or zero, thus generating BOW vectors.

2.5. Classifier Model Training and Evaluation – STEP 5-6

A supervised machine-learning approach is used to infer the extent to which significant vulnerabilities influence the determination of the severity of security risks. A ten-fold cross-validation process is applied for model validation. High model accuracy means that the input features have a significantly high descriptive power.

2.6. High-Impact Feature Selection – STEP 7

Our ultimate goal is to find the vulnerabilities that have a high impact on the severity of the security risk. The gain ratio method [3] enables us to find the most significant vulnerabilities. If a particular subset of vulnerability features shows the best performance in classification of high-severity class, the subset consisting of the vulnerabilities that are the most significant contributors to elevation of the severity of the security risk can be confirmed.

3. Conclusion

We have proposed a data-mining process to discover significant vulnerabilities that strongly affect the security of malicious-code-detected websites. Identification of high-impact vulnerabilities is very important for the development of proactive responses to risky websites, as discussed in this paper.

In the future, we will carry out an in-depth case study and discuss analytical experimental results. Besides, we plan to extend the current feature modeling by taking into account the environmental factors of risky websites.

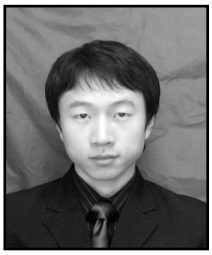
Acknowledgments

This research was supported by the Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (2012M3C4A7033345), and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A2009021).

References

[1] Current counter-measures and responses by CERTs: http://www.oecd.org/document/62/0,3746,en_2649_34255_38648830_1_1_1_1,00.html.
[2] J. Ma, L. K. Saul, S. Savage and G. M. Voelker, “Beyond Blacklists Learning to Detect Malicious Websites from Suspicious URLs”, the 15th ACM SIGKDD, **(2009)**.
[3] S. Shivaji, J. Whitehead, R. Akella and S. Kim, “Reducing Features to Improve Bug Prediction”, IEEE/ACM International Conference on Automated Software Engineering, **(2009)**.

Authors



Taek Lee is currently a Ph. D. candidate in Computer Science and Engineering at Korea University in Seoul, Korea. He received his MSc in Computer Science and Engineering at Korea University in 2006. His research interests include man-machine interaction, user behavior modeling in software systems, software defect prediction, information security, and information risk analysis.

Dohoon Kim received the B.S. in mathematics and dual degree in computer science & engineering at Korea University, Seoul, Korea, in 2005. He also received his M.S. and Ph. D. degrees in Computer Science & Engineering from the same University in 2007 and 2012. He is a senior researcher in the Information Technology Management Division from Agency for Defense Development. His current research interests are network security, risk management, software engineering, situation awareness, future internet and forecast engineering.

HyunCheol Jeong received his B.S. in Computer & Statistics from Seoul City University, Seoul, Korea in 1989, and his M.S. in Computer Science from KwangWoon University, Seoul, Korea in 1999. He is working for KISA(Korea Internet & Security Agency) since 1996. He was a senior technical member for KrCERT/CC and a director for Security R&D Team in KISA. Now, He is a director for IP address team in KISA. His current research interests are network security, digital forensic and incident response.

Hoh Peter In is a professor in Dept. of Computer Science at Korea University in Seoul, Korea. His primary research interests are requirements engineering, value-based software engineering, situation-aware middleware, and software security management. He created the WinWin requirements negotiation model for quality attributes as a team member. He has published over 100 research papers. He was an assistant professor at Texas A&M University. He received his Ph.D. from University of Southern California (USC) in 1998 and his B.S. and M.S. from Korea University in 1992 and 1994, respectively, all in computer science.