JAIST Repository
https://dspace.jaist.ac.jp/

Title	Drive-by-Download攻撃予測のための難読化JavaScriptの検知に関する研究
Author(s)	本田, 仁
Citation	
Issue Date	2016-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/13608
Rights	
Description	Supervisor:面 和成, 情報科学研究科, 修士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

A Study on Obfuscated JavaScript Detection for Drive-by-Download Attack Prediction

Jin Honda (1410039)

School of Information Science,
Japan Advanced Institute of Science and Technology

February, 2016

Keywords: Drive-by-Download, Obfuscated JavaScript, Machine Learning.

In recent years, cyber-attacks have increased with the popularization of the Internet. One of them is a Drive-by-Download attack (DbD attack), and it has become a serious threat. DbD attack causes malware download by exploiting vulnerabilities of Web browser and its plug-in without the users notice when a user accesses to a website that has been injected malicious codes by an attacker. In DbD attack, JavaScript often is exploited in the attack process. Most of JavaScript used in DbD attack is obfuscated in order to obscure the behavior, to evade detection mechanisms and to interfere with the analysis.

Therefore, obfuscated malicious JavaScript detection method using support vector machine (SVM) with character appearance frequency was proposed. This method focuss on the difference in characters that appear in normal JavaScript that is not obfuscated and obfuscated JavaScript. This method has the advantage that is simple and low cost in feature vector calculation. Also a grammatically incomplete JavaScript can also be the detection target because parsing and emulation are not necessary. In the evaluation of this method, cross-validation was carried out using D3M dataset of three years. However, the time series of datasets is not considered in cross-validation, a case that new data is learning data and old data is test data has occurred. It is unfair in the time series, thus this evaluation is considered to be not based on the detection in the real world.

In this research, we perform the experiment and evaluation based on the time series of the datasets, and show the effect of considering the time series by comparing these results with the results of the evaluation by cross-validation (not based on the time series). As a result, the evaluation score of SVM that is used in the existing method has deteriorated compared with Naive Bayes. Therefore, there is a possibility that Naive Bayes is effective at the detection in the real world.

In addition, we evaluate the character appearance frequency by bigram as a new feature, and examine the effectiveness. In evaluation by bigram, the dimension of the feature vector becomes enormous. We perform dimension reduction based on the degree of difference in the appearance frequency between benign and malicious data. The results of this evaluation experiment, the score is deteriorated compared with using unigram. It show that the character appearance frequency by bigram is not effective.

Copyright © 2016 by Jin Honda