Malicious URLs Filtering via Machine Learning Approach

李育杰
台灣科技大學資訊工程系

Abstract

Due to the advances of communication infrastructure and information technology, many internet services and Web applications have appeared. These bring a convenient life and also raise internet crimes at the same time. Malicious URLs have become a tool to activate the internet criminal activities such as malwares, spamming and phishing. There is a need to detecting the malicious URLs to protect the users get away from these malicious sites. Many detection methods have been proposed typically using the lexical and host-based features of the URLs or doing the content-based examination. However, it is impossible to examine a huge amount of URLs, say hundred millions that can be generated in one day and only have less than 0.01% malicious URLs. We propose a malicious URLs filter via machine learning approach. We generate two filtering models by using lexical features and dense features and then combine the filtering result. On-line learning algorithms are applied here not only for dealing with large scale datasets but also for fitting the characteristic of malicious URLs. Our proposed method is able to handle near two millions URLs less than five minutes. The filtering result can filter out 25% suspicious URLs that cover around 90% malicious URLs. It will reduce the burden for doing the content-based examination for URLs.