CITY UNIVERSITY OF HONG KONG

香港城市大學

ArrogateText: a Content-based Approach to Shanzhai Website Detection

ArrogateText:基於內容分析的山寨網站偵測方法

Submitted to

Department of Computer Science

電腦科學系

in Partial Fulfillment of the Requirements

for the Degree of Master of Philosophy

哲學碩士學位

by

LU Zhi

蘆 智

December 2010

二零一零年十二月

Abstract

Shanzhai website attack is a new kind of security threat that is harmful to the revenue and reputation of legitimate organizations. Despite this problem seems similar to the Phishing attack, which has been attracting a great many researchers for decades and many Anti-Phishing tools have been developed, Shanzhai website problem has its distinct features that make the current Anti-Phishing solutions infeasible to this problem.

Shanzhai websites usually use semantically similar content with the real websites, which is attacked by the Shanzhai websites, to deceive the visitors. In this thesis, a content-based Shanzhai website detection method, called ArrogateText, is proposed. The method first collects the candidate web pages from searching results of a given brand, in which the most strongly associated group of results are selected by extended TermCut. A matrix of these web pages' semantic relationships is constructed by analyzing the given web pages contents, where both the text semantic similarity and sentence structure similarity are considered. Finally, these web pages are ranked by multiplying the matrix and the ranking scores vector of the importance of the web pages. For a given website claiming a brand, if we find it is above certain position in the ranking list, we can determine it as the Shanzhai website. Otherwise, it should be treated as a legitimate website.

We design three experiments to evaluate the proposed method ArrogateText. As an early attempt to fight against the Shanzhai website problem, the proposed method achieves fairly good detection accuracy while remaining a relatively low false positive rate.

Table of Contents

List of Figures

List of Tables