International Journal of Innovative
Computing, Information and Control
Volume **6**, Number **5**, May **2010**

 ISSN 1349-4198
pp. **2301–2317**

A HARMLESS MECHANISM FOR STOPPING MALICIOUS CRAWLER

Jin-Cherng Lin[1] and Jan-Min Chen[1,2]

[1]Department of Computer Science and Engineer
Tatung University
40 Zhongshan North Road, 3rd Section Taipei 104, Taiwan
jclin@ttu.edu.tw

[2]Department of Information Management
Yu Da University
No.168, Hsueh-fu Rd, Miaoli County, Taiwan
ydjames@ydu.edu.tw

Received December 2008; revised June 2009

Abstract. *A number of automated software tools may be used to assist in identifying flaws in the Web, but unfortunately, these same tools may also be used to exploit these vulnerabilities. The web crawler, which is designed to discover all pages in a website, is a necessary part of these tools and its crawling rate can directly influence the effectiveness of other tools. The stopping of potentially malicious crawlers is an intuitive approach that can be applied to prevent vulnerabilities from being uncovered by attackers. Through our experiments and observations, we demonstrate that the crawling rate of a web crawler is dependent on the web techniques, programming languages and crawling techniques applied. In particular, it is demonstrated that a crawling screen in websites is necessary to effectively protect a website from enhanced crawlers, as we cannot rely on the hidden capabilities of web techniques or programming languages. This paper proposes a harmless scheme to accurately and automatically block specific crawlers according to the provided policy pertaining to any website. In comparison with other crawling screens, the implementation of our method in prototype shows that it not only increases security but also maintains the system's functionality and ease of use for normal visitors.*
Keywords: Crawler, Vulnerability scanner, Spider trap

1. **Introduction.** As the operation of web applications relies on attributes such as public availability, convenience of use and ease of location via a search engine, they present a very high risk as they are an attractive target for attackers. Due to the size and complexity of modern Web-based applications, the process of discovering existing web vulnerabilities manually is an extremely time consuming task. Thus, in recent years, some tools have been designed to automate this process. Nevertheless, though these tools assist in finding flaws, they may also be misused to exploit the vulnerabilities of web applications. Upon inspection of the processes involved in these tools, we discover that the component central to exploiting vulnerabilities is the web crawler, which is designed to discover all pages in a website. Thus, the blocking of all crawlers would be the most direct and effective way to prevent the uncovering and exploitation of vulnerabilities.

However, there are obstacles to such a method, which necessitate making crawler blocking a discriminatory mechanism. In particular, the fact that some recently devised tools have embedded crawlers to facilitate performing a number of useful tasks such as statistical analysis, maintenance of the hypertext structure, implementation of Web mirroring