Towards More Effective Web Crawler Detection Mechanisms

Anália Lourenço analia@di.uminho.pt
Orlando Belo obelo@di.uminho.pt

Departamento de Informática, Escola de Engenharia, Universidade do Minho
Campus de Gualtar
4710-057 Braga
PORTUGAL

Abstract. Web crawlers are one of today's most active Web communities. The steady growth of Web contents and their dynamism impelled the creation of these automatic information retrieval programs. Users rely on the updated indices of search engines and directories in order to be succeeded in their searches. Search engines and directories require exhaustive crawling work to maintain and update their indices to the increasingly time-sensitive Web content. The need for Web crawlers is obvious and unavoidable. However, as happened with many other programs, Web crawlers started to have devious applying due to careless and bad intentioned programming. The absence of strict, effective regulation leaves room to dangerous situations and ultimately, to fraud. Understanding the nature and primary characteristics of Web crawlers is an essential step to analyse their impact on Web sites and Web itself. Another crucial step is to deploy detection and filtering mechanisms capable of efficiently delimiting the access of Web crawlers to Web sites. Regarding this, navigation pattern mining is issued as a way of interpreting Web crawlers' behaviour, delivering the required information for the detection heuristics. This paper analyse the impact of Web crawlers activities on the Web through the understanding of their nature and primary characteristics, giving particularly attention to the deployment of detection and filtering mechanisms capable of efficiently delimiting the access of Web crawlers to Web sites.

Keywords: Search Engines, Web Crawlers, Clickstream Processing, and Navigation Pattern Mining.