International Journal of Computer Science
and Engineering (IJCSE)
ISSN(P): 2278-9960; ISSN(E): 2278-9979
Vol. 6, Issue 1, Dec - Jan 2017; 1-34

International Academy of Science,
Engineering and Technology
Connecting Researchers; Nurturing Innovations

FOCUSED WEB CRAWLER DEVELOPMENT CHALLENGES: ECCRAWLER

FURKAN GÖZÜKARA & SELMA AYŞE ÖZEL

Department of Computer Engineering, Faculty of Engineering and Architecture,

Çukurova University, Balcalı, Sarıçam, Adana, Turkey

ABSTRACT

Nowadays, the importance of focused web crawlers is more than any time before. As the web has become massive and spam my, it is now essential to have focused web crawlers that can crawl only the targeted websites and obtain the necessary information. Instead of relying on the available public general web crawlers, today, developing a focused web crawler for the targeted web pages is preferred to increase success of information retrieval. In this paper, the challenges encountered and the proposed solutions to attempt these problems are presented, while developing an original hand-crafted, full scale, robust and effective focused web crawler for E-commerce sites, named as EcCrawler, which is developed in C# programming language by using .NET 4.5 framework and MS-SQL Server 2014 database management system. Most of the crawling challenges have been discussed before in the literature, however in this paper, practical implementation and .NET framework based solutions that includes thread pool initialization, exception handling, task parallelism, HTTP compression, duplicate web page resolution, number of concurrent connections to the same host, database communication, resource sharing between threads, etc. are presented and the proposed solutions are empirically evaluated. The experimental evaluation shows that applying the proposed solutions improve EcCrawler's crawling speed over 400% and UI responsiveness over 100%. The proposed solutions may be applicable to any software that is developed by using .NET framework.

KEYWORDS: NET Framework, Performance Tuning, Application Development, Multithreading, Web Crawling