Abstract

The Web is a context in which traditional Information Retrieval methods are challenged. Given the volume of the Web and its speed of change, the coverage of modern web search engines is relatively small. Search engines attempt to crawl the web exhaustively with crawler for new pages, and to keep track of changes made to pages visited earlier. The centralized design of
Design and Implementation of Scalable, Fully Distributed Web Crawler for a Web Search Engine

crawlers introduces limitations in the design of search engines. It has been recognized that as the size of the web grows, it is imperative to parallelize the crawling process. Contents other than standard documents (Multimedia content and Databases etc) also makes searching harder since these contents are not visible to the traditional crawlers. Most of the sites stores and retrieves data from backend databases which are not accessible to the crawlers. This results in the problem of hidden web. This paper proposes and implements DCrawler, a scalable, fully distributed web crawler. The main features of this crawler are platform independence, decentralization of tasks, a very effective assignment function for partitioning the domain to crawl, and the ability to cooperate with web servers. By improving the cooperation between web server and crawler, the most recent and updates results can be obtained from the search engine. A new model and architecture for a Web crawler that tightly integrates the crawler with the rest of the search engine is designed first. The development and implementation are discussed in detail. Simple tests with distributed web crawlers successfully show that the DCrawler performs better then traditional centralized crawlers. The mutual performance gain increases as more crawlers are added.

Reference


Index Terms

Computer Science Information Retrieval

Key words

web search engines crawling software architecture