Abstract

In today's information era, Web becomes one of the most powerful and fastest means of communication and interaction among human beings. Search engines as Web based applications traverse the Web automatically and receive the set of existing fresh and up-to-date documents. The process of receiving, storing, categorizing and indexing is done automatically based on partial smart algorithms. Although many facts about the structure of these applications remains hidden as commercial secrets, the literature tries to find the best approaches for each modules in the structure of search engines. Due to the limited time of today's Web surfers, providing the most related and freshest documents to them is the most significant challenge for search engines. To do so, every module in search engine architecture should be designed as smart as possible to yield not only the most related documents but also to act in a timely manner. Among these modules is the sensitive part of crawler. One of the open issues in optimization of search engines' performance is to reconfigure crawling policy in a way that it follows the most promising out-links that carries the content related to the source page. Crawler module has the responsibility to fetch pages for ranking modules. If higher quality pages with less content drift are indexed by the crawlers, the ranking module will perform faster. According to the graph structure of the Web, the way of traversing the Web is based on the literature on graph search methods. This paper
experimentally employs different graph search methods and different combinations of them by issuing some queries to Google engine to measure the quality of received pages with fixing the factor of graph depth to identify the best method with reasonable time and space complexity to be employed in crawler section in search engine architecture.

References

- Esmaeeli, m. tavakoli,m, hashemi majd, s, 2013. "The Web crawler" APA professional laboratory in context of information and communication technology security, document number, APA_FUM_W_WEB_0111, pp. 5-28, bahman.
Page Quality Optimization in Crawler’s Queue through Employing Graph Traversal Algorithms


Index Terms

Computer Science
Algorithms

Keywords