Abstract

The size of the WWW is increasing rapidly and its nature is dynamic, building an efficient search mechanism is very necessary. A vast number of pages continually being added every day, so fetching information about a special-topic is gaining importance, which poses exceptional scaling challenges for general-purpose crawlers and search engines. This paper describes a web crawling approach based on best first search. Instead of collecting and indexing all available web documents to be able to answer all possible queries, a focused crawler choose the links that are likely to be most relevant for the crawl, and avoids irrelevant links of the document. This leads to significant savings in hardware as well as network resources and also helps keep the crawl more up-to-date. To accomplish such goal-directed crawling, select top most K relevant documents for a given query and then expand the most promising link chosen according to link score, to circumvent irrelevant regions of the web.
Focused Crawler based on Efficient Page Rank Algorithm

- Bing Liu, "Web Content Mining" the 14th international world wide web conference
- M. Gori, M. Maggini, and F. Scarselli, "http://nautilus.dii.unisi.it."
- Page Rank form Wikipedia, the free encyclopedia http://en.wikipedia.org/wiki/PageRank/

Index Terms

Computer Science
Web Services
Keywords
Focused web crawler  TF-IDF  Relevancy calculation  Page Rank.