Abstract

The immense growing dimension of the World Wide Web induces many obstacles for all-purpose single-process crawlers including the presence of some incorrect answers among search results and the scaling drawbacks. As a result, more enhanced heuristics are needed to provide more accurate search outcomes in an appropriate timely manner. Regarding the fact that employing link dependent Web page importance metrics within a parallel crawler yields a considerable overhead on the overall searching system, and also because such a metric is not able to cover the authorized Web content in dark net and authorized fresh pages, therefore employing these metrics is not an absolute solution within search engines' architecture. This paper proposes the application of a link independent Web page importance metric to govern the priority rule within the crawl frontier through proposing a modest weighted architecture for a focused structured parallel Web crawler (CFP crawler) in which the credit assignment to URLs
in crawl frontier is done according to a clickstream-based prioritizing algorithm.

Reference


Index Terms

Computer Science
Internet Technology

Key words
Clickstream analysis
Focused crawlers
Parallel crawlers
Web data management
Web page Importance metrics