Abstract

The rapid growth of the World Wide Web (WWW) poses unprecedented scaling challenges for general-purpose crawlers. Crawlers are software which can traverse the internet and retrieve web pages by hyperlinks. The focused crawler of a special-purpose search engine aims to selectively seek out pages that are relevant to a pre-defined set of topics, rather than to exploit all regions of the Web. Focused crawler is developed to collect relevant web pages of interested topics from the Internet. Maintaining currency of search engine indices by exhaustive crawling is rapidly becoming impossible due to the increasing size of the web. Focused crawlers aim to search only the subset of the web related to a specific topic, and offer a potential solution to the problem. In our proposed approach, we calculate the link score based on average relevancy score of parent pages (because we know that the parent page is always related to child page which means that for detailed information any author prefers the child page) and division score (means how many topic keywords belong to division in which particular link belongs). After finding out link score, we compare the link score with some threshold value. If link score is greater than or equal to threshold value, then it is relevant link. Otherwise, it is discarded. Focused crawler first fetches that link which has greater value compared to all link scores and threshold.
Reference


Index Terms

Computer Science Internet Applications

Key words

Focused crawler Crawler
Division score
Link score