Abstract

The Web provides us with a huge and endless resource for information. But, the rapidly growing size of the Web poses great challenge for general purpose crawlers and search engines. It is impossible for any search engine to index the whole Web. Focused crawler collects domain relevant pages from the Web by avoiding the irrelevant portion of the Web. Focused crawler can help the search engine to index all documents present on the Web related to a specific domain which in turn provides the search engine’s users complete and up-to-date contents. In this paper we present a focused crawler capable of learning from the previous crawl results to collect the relevant documents. Crawling results for three consecutive learning phases are shown. Results indicate significant improvement in terms of relevancy to the focused domain.

References

World Wide Web, Hong Kong, pp. 96-105.
- Ari Pirkola (2007), "Focused Crawling: A Means to Acquire Biological Data from the Web," VLDB &apos;07, Vienna, Austria.

Index Terms
Keywords
Web  Internet  Retrieval  Focused Web Crawler  Search Engine Etc