Abstract

The web is a huge and highly dynamic environment which is growing exponentially in content and developing fast in structure. No search engine can cover the whole web, but it has to focus on the most valuable pages for crawling. Many methods have been developed based on link and text analysis for retrieving the pages. In this paper, an algorithm based on link, text, logarithmic distance and probabilistic measure is presented to find the relevancy of the web pages. Here, the most relevant pages are retrieved. It has been proved experimentally that this method provides more number of relevant pages.

References

- Sandeep Sharma, Mr. Ravinder Kumar, Web-Crawling Approaches in Search Engines, June 2008.
Index Terms

Computer Science
Information Sciences

Keywords
Search engine  Logarithmic distance  Probabilistic measure