Abstract

Vertical search engines use focused crawler as their key component and develop some specific algorithms to select web pages relevant to some pre-defined set of topics. 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. Maintaining currency of search engine indices by exhaustive crawling is rapidly becoming impossible due to the increasing size of the web. Focused crawler aims to search only the subset of the web related to a specific topic, and offer a potential solution to the problem. A focused crawler is an agent that targets a particular topic and visits and gathers only a relevant, narrow web segment while trying not to waste resources on irrelevant material. As the crawler is only a computer program, it cannot determine how relevant a web page is. The major problem is how to retrieve the maximal set of relevant and quality page. In our proposed approach, we classify the unvisited URL based on visited URLs attribute score, i.e., unvisited URLs are relevant to topics or not, and then decide based on seed page attribute score. Based on score, we put “Yes” or “No” values in the table. URLs attributes are: it's Anchor text relevancy, its
Unvisited URL Relevancy Calculation in Focused Crawling Based on Naïve Bayesian Classification

description in Google search engine and calculates the similarity score of description with topic keywords, cohesive text similarity with topic keywords and Relevancy score of its parent pages. Relevancy score is calculated based on vector space model. Classification is done by Naïve Bayesian classification methods.

Reference


Index Terms

Computer Science  Knowledge Discovery

Key words

Crawler  Focused crawler
Vector

Naïve Bayesian classification methods