Abstract

The constraint of a web crawler that downloads only relevant pages is still a major challenge in the field of information retrieval systems. Rather than visiting all the web pages, a focused
crawler visits only the section of the web that contains relevant pages, and at the same time, tries to skip irrelevant sections. Existing ontology based web crawlers estimate the semantic content of the URL based on a domain dependent ontology, which in turn supports the methods used for prioritizing the URL queue. The crawler maintains a queue of URLs it has seen during the crawl at each level, and then selects from this queue, the next URL to visit based on the conceptual rank of the page at that level obtained from domain ontology. However in this work we represent the topic as an overall conceptual vector, obtained by combining concept vectors of individual pages associated with seed URLs. The conceptual rank is based on comparison between conceptual vectors at each depth, across depths and between the overall topics indicating seed concept vector.

Reference

- Debashis Hati, Amritesh kumar, 2010. An approach for identifying URLs based on Division score and link score in focused crawler, International journal of computer applications, Volume 2 – No.3.

Index Terms
Concept based Focused Crawling using Ontology

Computer Science

Information Retrieval

Key words
Focused crawler  Ontology  Conceptual vector