Abstract

The Internet has billions of web pages and these web pages are attached to each other using URL (Uniform Resource Allocation). Web crawler is a main module of Search engine that gathers these documents from WWW. Most of the web pages present on Internet are active and changes periodically. Thus, Crawler is required to update these web pages to update database of search engine. In this paper, priority based semantic web crawling algorithm has been proposed. Ontology is used to get semantics of web page during crawling process. Algorithm starts with initial seed URL. The web page at given URL is downloaded from Internet and semantic score is calculated with given topic. The semantic score of unvisited URL is calculated using its Anchor text semantic similarity score, semantic similarity score of web page of unvisited URL with given topic and semantic score of its parent pages. Priority queue is used to store URL and its semantic score instead of simple queue. So, every time priority queue returns higher priority URL to crawl next. The overall performance gain over simple crawler is 88%, over focused crawling is 28% and priority based focused crawler is 6%.

References


Index Terms

Computer Science

Information Sciences
Keywords

Priority ontology Semantic similarity downloader search engine