Abstract

The World Wide Web is a huge source of hyperlinked information contained in hypertext documents. Search engines use web crawlers to collect these documents from web for the purpose of storage and indexing. However, many of these documents contain dynamic information which gets changed on daily, weekly, monthly or yearly basis and hence we need to refresh the search engine side storage so that latest information is made available to the user. An incremental crawler visits the web repeatedly after a specific interval for updating its collection. In this paper to regulate the revisiting frequency a novel mechanism and a novel architecture for incremental crawler is being proposed.

Reference


Index Terms

Computer Science

Internet Applications
Design of a Priority Based Frequency Regulated Incremental Crawler

Key words

web search engine
incremental crawler
dynamic
information
crawl workers