Distributed Web Crawler

Implementation of a distributed Web Crawler

The Web is a context in which traditional Information Retrieval methods are challenged. Given the volume of the Web and its speed of change, the coverage of modern web search engines is relatively small. Search engines attempt to crawl the web exhaustively with crawler for new pages, and to keep track of changes made to pages visited earlier. This results in the problem of hidden web. This paper proposes and implements Distributed Web Crawler, a scalable, fully distributed web crawler. The main features of this crawler are platform independence, decentralization of tasks, a very effective assignment function for partitioning the domain to crawl, and the ability to cooperate with web servers.

Algorithm:

1) Download the corresponding document and its HTTP header using a GET request.

2) Parse and extract all links contained in the document.

3) Proceed with the canonicalization and normalization of URLs.

4) Apply a URL filter, keeping only URLs that match the user-supplied configuration, file type, and domain specific URLs.

As a result, for each URL each worker returns its HTML document, HTTP header details, and the list of discovered out-links. The URL-seen function checks if the URL has already been processed or queued before being added it to the frontier.

Steps or commands to run the code:

python3 controller.py

Crawler.py file:

In crawler file a class is initialized which crawls through the base URL. The base URL is given as input to the object of this class. It also writes the logs, dumps and error logs into respective files.

Controller.py file:

In this file an object is initialized for the class defined in crawler file. This file must be executed to get the results. It invokes all the required functions and performs crawling of the base URL.

Error_log.txt : The error logs are written into this file.

Logs.txt: The crawled URLs are written into this file.

Dumps.txt: The duplicate URLs are written into this file.

Conclusion:

The Web crawler can efficiently download several million pages per day and can be used for web search and web characterization. Several web portals have been downloaded and analyzed using Distributed Crawler, extracting statistics on HTML pages, multimedia files, Web sites and link structure.