DESIGN AND IMPLEMENTATION TECHNIQUE FOR PARALLEL CRAWLER

By Dr.Mukesh Rawat

WWW is works on the internet client-server network topology, it assist in providing information in the form of hypertext (in html form). Crawler is one of the important module of search engine that crawls the interconnected web pages of a particular website and helps search engine to identify relevant document web pages according to the user query. The article explains a system that identifies link in the document in the form of text or images.

As huge number of html web pages are available on World Wide Web so, a search engine uses web crawlers to fetch the web pages from WWW.

Web crawler/spider fetches the hyperlinks presented in a specific web page and then follow each collected hyperlink, download web pages and stored in the search engine database. There is a module named indexer, which indexes the collected web pages by using indexing techniques such as "Inverted Indexing". The indexing of web pages helps the search engine to retrieve list of similar web pages available in the posting of a particular term given by the user. As shown in figure 1.

As today 5.47 billion web **pages** available in world wide web and the counting of these web pages are increasing enormously day by day. So, the availability of such large number of web pages requires a effective mechanism to crawl these web pages.

There are multiple web crawlers available such as form focused web crawler, which extract out the web pages containing search interfaces. As the number of web pages are large so, the load on crawler increases, to reduce this load of crawling on a single crawler parallel crawler comes into picture where the crawling process distributed among different parallel crawlers.

Online digital library uses many parallel crawling process to mine the world wide web. Each web crawler works according to one thread model which refers a collection of initiating URLs and extract out pages in parallel. The hyperlinks fetched from each web page stored in a separate data structure, it may be vector, array list etc.

A crawler is a program that starts from a seed url and fetches interconnected web pages.

It is very time consuming and not significant engage a single web crawler to fetch the entire or important part of web takes much more time and very tedious to finish. Therefore, most of the search engines running many copies of parallel web crawlers/spiders to accomplish the task of web crawling, to achieve maximum .

There is some of the disadvantage of parallel crawling, as different parallel crawlers download the same page results in redundant crawling. But parallel/simultaneous web crawling has many benefits in comparison to single crawler such as it can be changed in scale or size, distribution of work among network nodes and reduction.

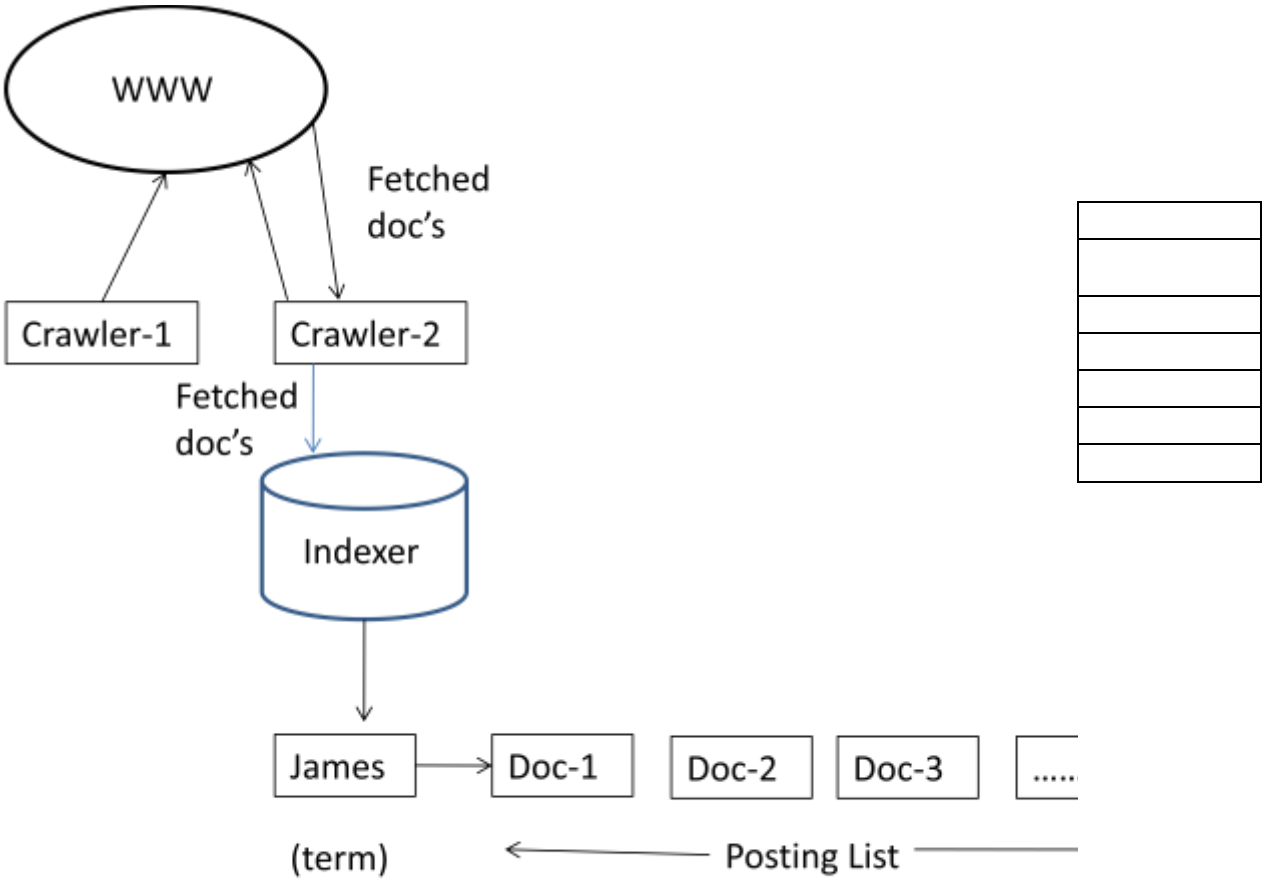


Fig 1: Creation of posting List by indexer

WORKING OF A PARALLEL CRAWLER

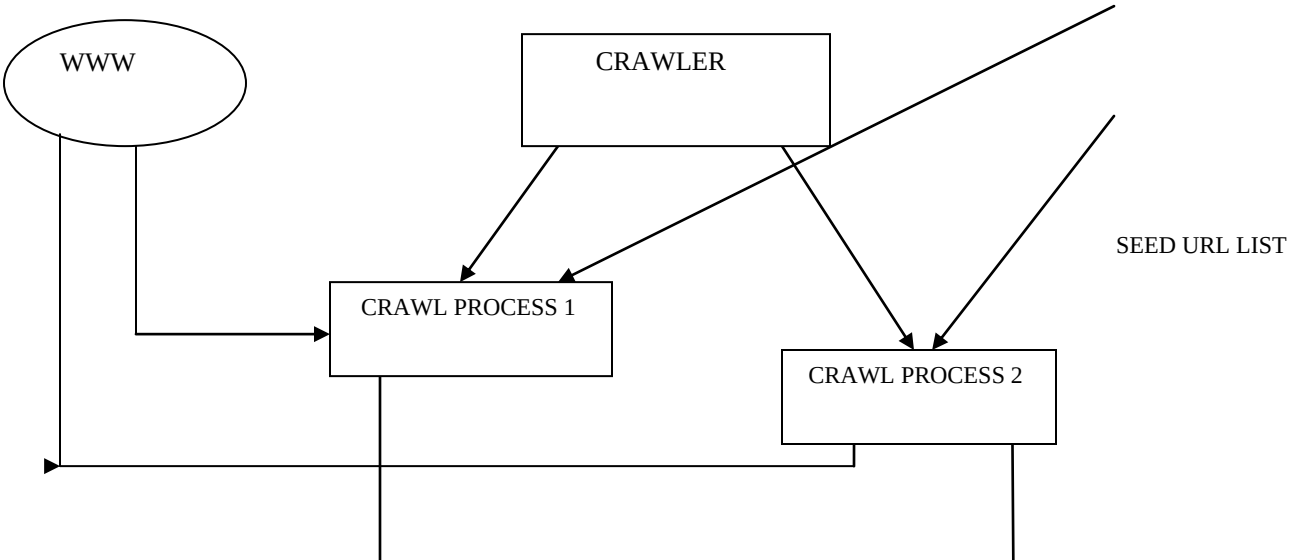
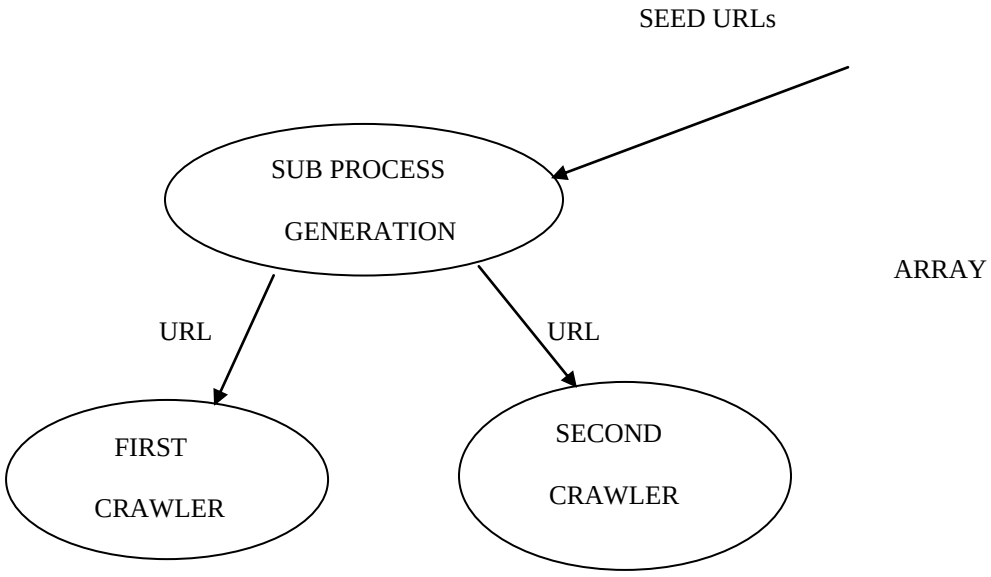


Fig 2: General Architecture for Parallel Crawler

DETAILED DESCRIPTION OF NTERCONNECTED MODULES OF PARALLEL CRAWLER



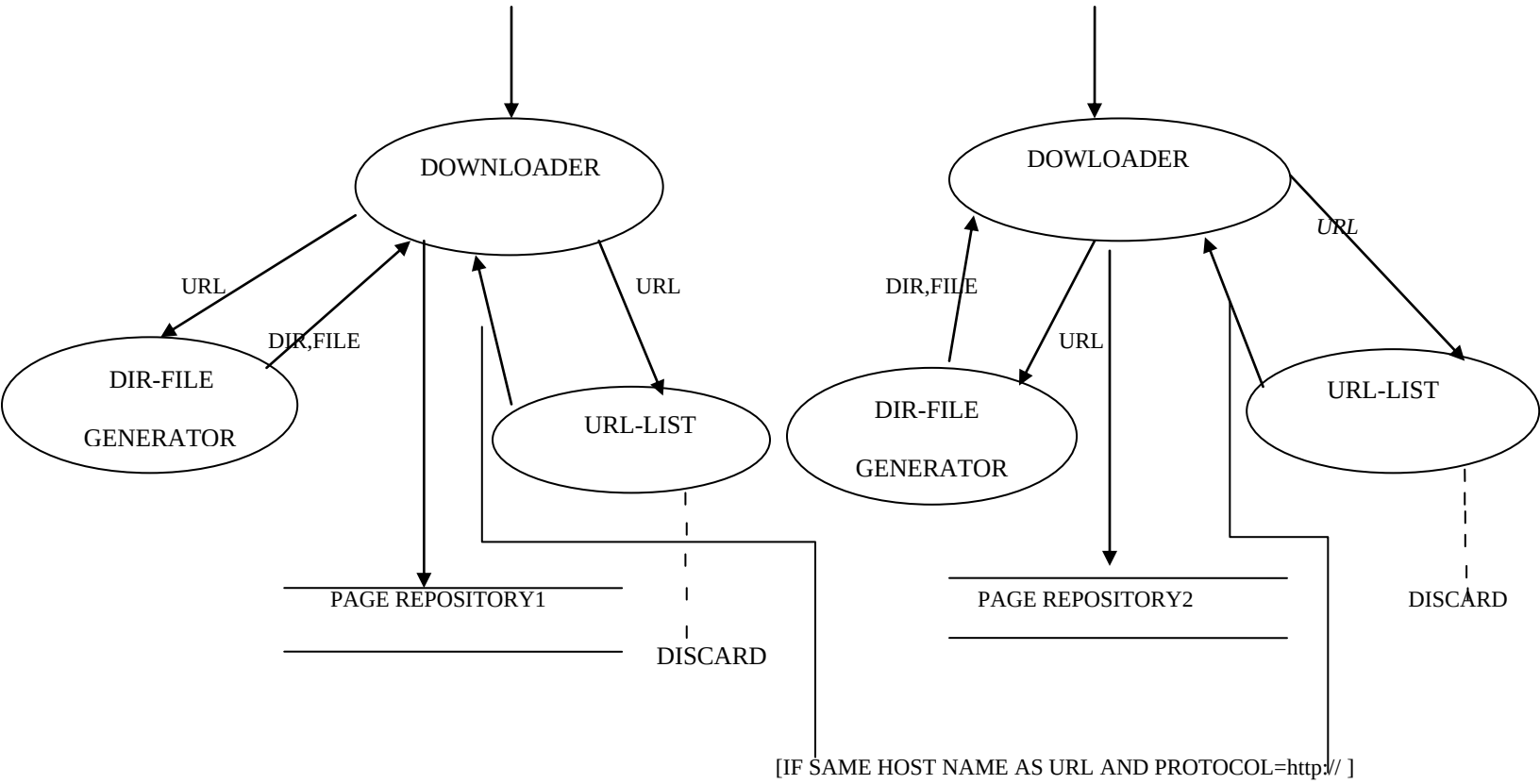


Fig 3: Working method of each parallel crawler

Module -Description

a: SUB PROCESS GENERATION:

The Sub-Process-Generation method creates multiple parallel crawler according to the list of urls given in Seed URLs list. Each crawler process work as a single thread and extract pages linked in the picked seed URL .All the crawler modules fetch pages simultaneously. The hyper-links are fetched from downloaded html documents and put in separate list (any dynamic data structure).

b: DOWNLOADER:

Downloader extracts out the hyperlinks available in the target html document and stores them in one of the dynamic data structure such as vector, arraylist etc. Then recursively pick the

hyperlink from the link storage. This process going till the link storage becomes empty. The working of "Downloader" is described in the fig 4.

c: DIR-FILE GENERATOR

Dir-File generator creates the same directory and file structure and hierarchy in the system from where the parallel crawler is working. So, that the proper distribution of files and directories are maintained.

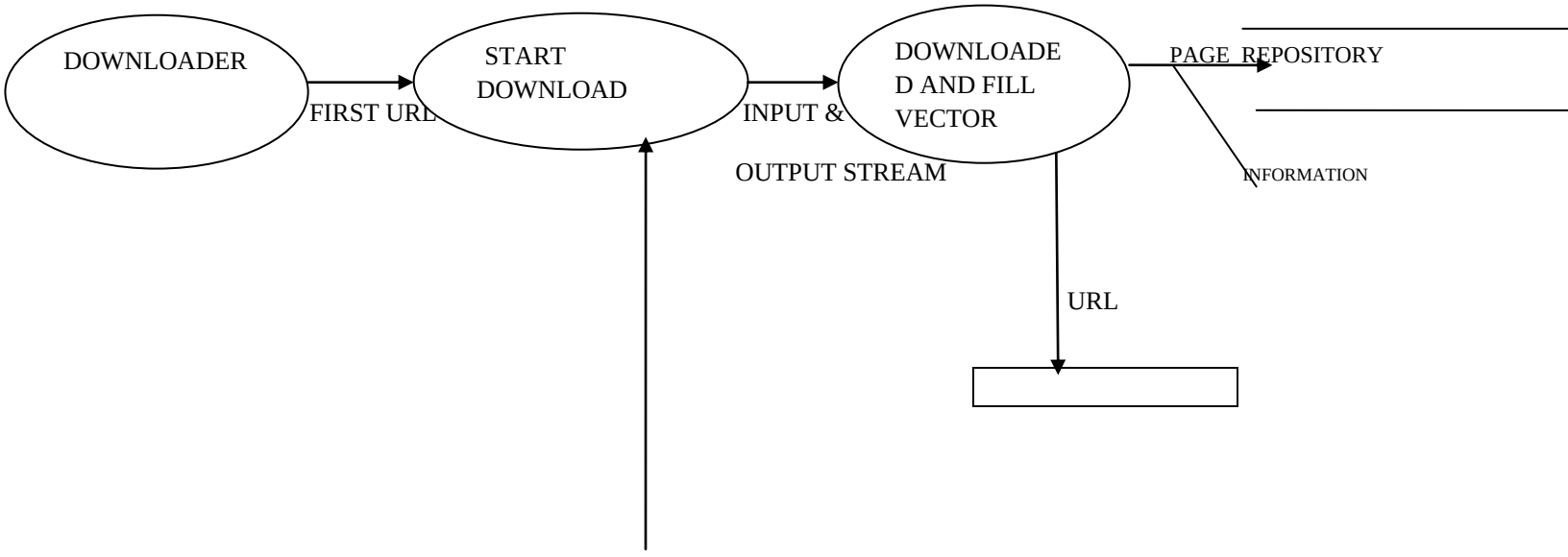
d: URL LIST

This module checks whether the hyperlink belongs to the same domain as specified in the seed url used by parallel crawler. For example , if seed url given as http://www.abc.com/jstl/index.com and if the parallel crawler receives http://www.jango.com/py/index.com. As the domain name of both the hyperlinks are different in this case crawling module discard such links. As it increases crawling time.

e: PAGE REPOSITORY

Page Repository, stored relevant pages that are downloaded by the downloader of respective parallel crawler.

Working of downloader



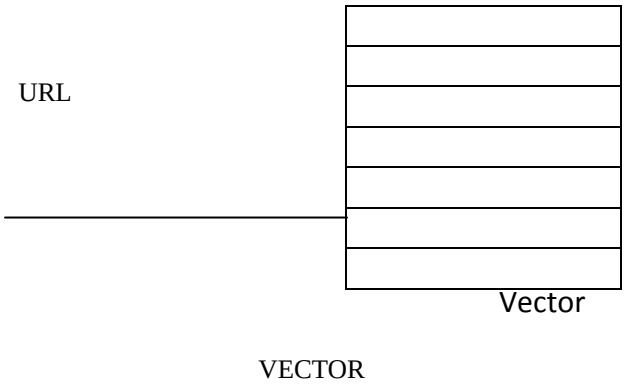


Fig 4: Process used in downloader module

The crawling process may be going long enough as fetch pages contains url's of other pages, the crawling process limited by applying domain restriction or fix the number of hyperlinks fetched.