Project name: Ayudika

Subject: Building a web search engine

Web Search Engine:

Web Search Engine is a software that is used to search an information form world wide web.

World Wide Web collection of websites or web pages stored in web servers and connected to local computers through the internet. These websites contain different types of informations. Users can access the content of the sites from any part of the world over the internet using their devices such as computers, laptops, cell phones, etc.

Web Search Engine Architecture:

1. Web Crawler
2. Indexer
3. Quarry Processor

Special Features of Web Search Engine:

1. Indexing large no of documents
2. Prevent hacks and spams
3. Filter bubbles(ie, local search results)
4. HTTPS redirects
5. Understanding users quarry
6. Auto complete

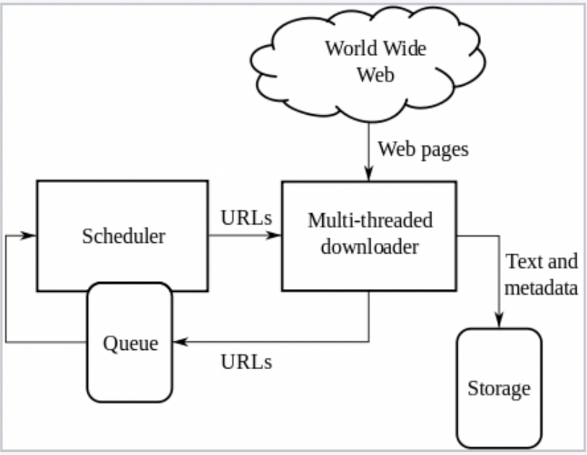
Working of Web Search Engine:

There are some Important parts of Search Engine -

1. Web Crawlar :

Web Crawler is a software which is also called called a Spider or Spiderbot.

- It is used to finds Informations ie, documents from World Wide Web.
- Crawler starts crawling form seed url (generally we use wikipedia as a seed url because of no spam and high page rating.
- Also Crawler compresses the data and urls
- For avoiding loop use maximum 2 redirects



Architecture of a Web crawler

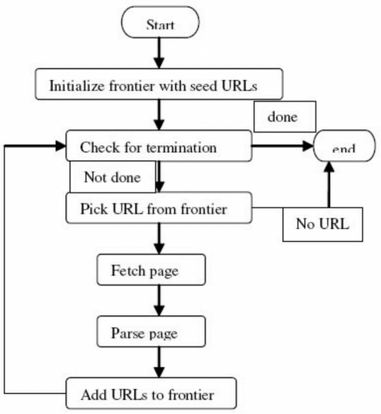


Fig. 2: Flow chart representing a web crawler

Some Web Crawling algorithms are:

- Breadth First Search Algorithm
- Depth First Search Algorithm
- Page Rank Algorithm
- Focused Crawling Algorithm

Some Important links:
https://www.computerscijournal.org/vol7no1-2/an-algorithm-for-effective-web-crawling-mechanism-of-a-search-engine/

2. Parser:

Parser is a software which is responsible for scanning the documents and links int the documents provided by Web Crawler.

- Mainly parser retrieves HTML codes
- Parser decompresses data, extensions, tital tag, meta tag, description tag, h1 tag and **gives to Indexer.**
- Parser also finds anchor tags and gives to web frontier and then it is given to Crawler again.
- 301 error ans 302 error

  301 error: Says permanent change in url ie, server down

  302 error: Says temporary changes in url

3. Indexer:

Indexer is a part of Search Engine that arranges the documents for quarry processing.

4. Quarry processor:

There are some major responsibilities of Quarry Processor like-

1. Checking spellings.
2. Auto Completion
3. Find informations4.

4.Text Checking

Flow of software:

Main.html -> Crawlar -> Compressor -> Parser -> Decompressor -> Parser -> webFrontier(<a herd = “ “ > <\a>) -> Crawlar