Web Crawler

All of us use search engines almost daily. When most of us talk about search engines, we really mean the World Wide Web search engines. A very superficial overview of a search engine suggests that a user enters a *search term* and the search engine gives a list of all relevant resources related to that *search term*. But, to provide the user with a list of resources the search engine should know that they exist on the Internet. Search engines do not magically know what websites exist on the Internet. Therefore, this is exactly where the role of web crawlers comes into the picture.

- What is a web crawler?

A web crawler is a bot that downloads and indexes content from all over the Internet. The aim of such a bot is to get a brief idea about or know what every webpage on the Internet is about so that retrieving the information becomes easy when needed. The web crawler is like a librarian organizing the books in a library making a catalog of those books so that it becomes easier for the reader to find a particular book. To categorize a book, the librarian needs to read its topic, summary, and some part of the content if required.

- What is the main process of web crawler programs?

They **crawl the webpages at those URLs first. As they crawl those webpages, they will find hyperlinks to other URLs, and they add those to the list of pages to crawl next**. Given the vast number of webpages on the Internet that could be indexed for search, this process could go on almost indefinitely.

It is very difficult to know how many webpages exist on the Internet in total. A web crawler starts with a certain number of known URLs and as it crawls that webpage, it finds links to other webpages. A web crawler follows certain policies to decide what to crawl and how frequently to crawl.
Which webpages to crawl first is also decided by considering some parameters. For instance, webpages with a lot of visitors are a good option to start with, and that a search engine has it indexed.

That's why we should specify the depth. (how many URLs to crawl from one web page to another)

- What is web crawler example?

For example, **Google has its main crawler, Googlebot, which encompasses mobile and desktop crawling**. But there are also several additional bots for Google, like Googlebot Images, Googlebot Videos, Googlebot News, and AdsBot.

- What type of agent is web crawler?

A bot. A Web crawler is one type of **bot, or software agent**.

These are the following steps to create a web crawler:

1. In the first step, we first pick a URL from the specified URL list.
2. Fetch the HTML code of that URL.
3. Get the links to the other URLs by parsing the HTML code.
4. Check whether the URL is already crawled before or not. We also check whether we have seen the same content before or not. If both conditions don't match, we add them to the **index.**
5. For each extracted URL, verify that whether they agree to be checked (robots.txt, crawling frequency)

Web crawler algorithm

```
public void getPageLinks(String URL) {
        //4. Check if you have already crawled the URLs
        //(we are intentionally not checking for duplicate content in this example)
        if (!links.contains(URL)) {
            try {
                //4. (i) If not add it to the index
                if (links.add(URL)) {
                    System.out.println(URL);
                }

                //2. Fetch the HTML code
                Document document = Jsoup.connect(URL).get();//jsoup jar to extract web
data
                //3. Parse the HTML to extract links to other URLs
                Elements linksOnPage = document.select("a[href]");

                //5. For each extracted URL... go back to Step 4.
                for (Element page : linksOnPage) {
                    getPageLinks(page.attr("abs:href"));
                }
            } catch (IOException e) {
                System.err.println("For '" + URL + "': " + e.getMessage());
            }
        }
    }
```

https://subscription.packtpub.com/book/big-data-and-business-intelligence/9781787122536/1/ch01lvl1sec20/extracting-web-data-from-a-url-using-jsoup