Search Engine

Search Engine

A **search engine** is a type of computer software used to search data in the form of text or a database for specified information.

Search engines normally consist of spiders (also known as bots) which roam the web searching for links and keywords. They send collected data back to the indexing software which categorizes and adds the links to databases with their related keywords. When you specify a search term the engine does not scan the whole web but extracts related links from the database.

Please take note that this is **not** a simple process. Search engines literally scan through millions of pages in its database. Once this has taken place all the results are put together in order of relevancy. Remember also not to get a search engine and directory mixed up. Yes they are used interchangeably, but they do in fact perform two different tasks!

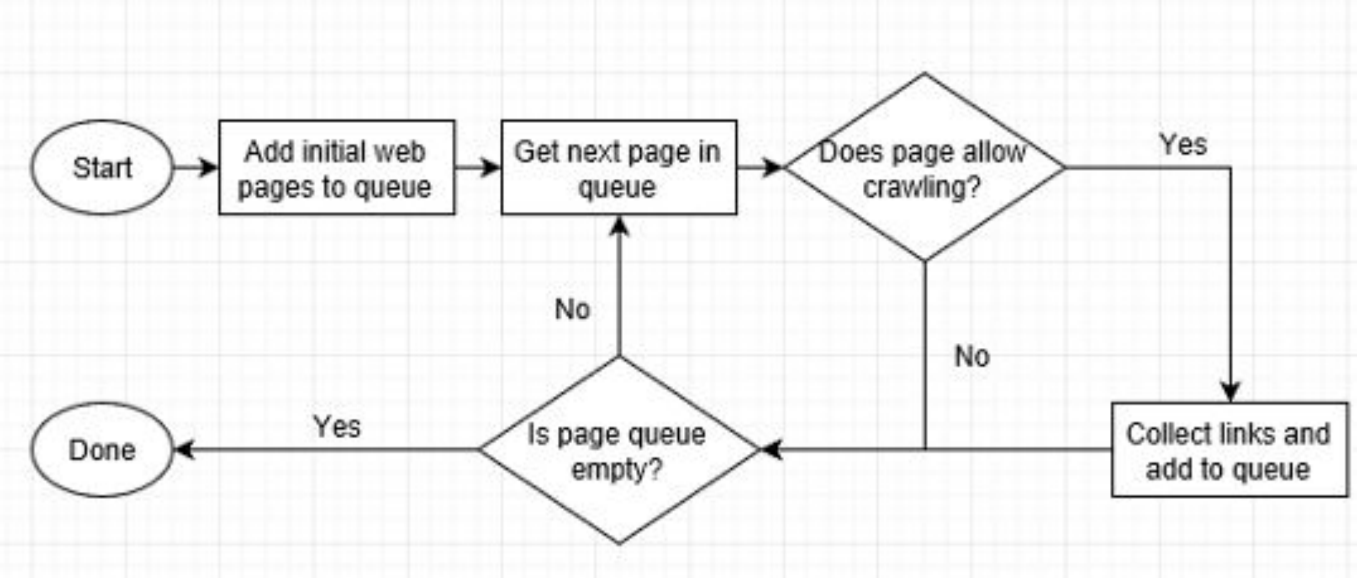
Before 1993 the term search engine never existed. From then until now that has changed drastically, and almost everyone knows what it is. Since the Internet is used by millions of Americans daily, a search engine sees a lot of visitors especially ones such as Google and Yahoo. Almost all of us use one of the two if we have the Internet. By simply typing words into the engine, we get several results which gives us a list of sites. (Seigel)

Usually a search engine sends out a spider which fetches as many documents as possible. An index is then created by what is called an indexer that reads the documents and creates it. Only meaningful results are created for each query though, a process called proprietary algorithm.

Crawling

Crawling is where it all begins: the acquisition of data about a website.

This involves scanning sites and collecting details about each page: titles, images, keywords, other linked pages, etc. Different crawlers may also look for different details, like page layouts, where advertisements are placed, whether links are crammed in, etc.


Start
Add initial web pages to queue
Get next page in queue
Does page allow crawling?
Yes
No
Collect links and add to queue
Is page queue empty?
No
Yes
Done

But how is a website crawled?

Crawling is where it all begins: the acquisition of data about a website.

This involves scanning sites and collecting details about each page: titles, images, keywords, other linked pages, etc.

Different crawlers may also look for different details, like page layouts, where advertisements are placed, whether links are crammed in, etc.

But how is a website crawled? An automated bot (called a “spider”) visits page after page as quickly as possible, using page links to find where to go next. Even in the earliest days, Google’s spiders could read several hundred pages per second. Nowadays, it’s in the thousands.

Indexing

Indexing is when the data from a crawl is processed and placed in a database.

Imagine making a list of all the books you own, their publishers, their authors, their genres, their page counts, etc. Crawling is when you comb through each book while indexing is when you log them to your list.

Now imagine it's not just a room full of books, but every library in the world. That's a small-scale version of what Google does, who stores all of this data in vast data centers with thousands of petabytes worth of drive.