**WEB CRAWLER in SEO (Search Engine Optimization)**

Introduction to Web Crawlers –

A search engine crawler is a program or automated script that browses the World Wide Web in a methodical manner in order to provide up to date data to the particular search engine. While search engine crawlers go by many different names, such as web spiders and automatic indexers, the job of the search engine crawler is still the same.

The process of web crawling involves a set of website URLs that need to be visited, called seeds, and then the search engine crawler visits each web page and identifies all the hyperlinks on the page, adding them to the list of places to crawl. URLs from this list are re-visited occasionally according to the policies in place for the search engine. The policies of the search engine can be different for each search engine, and may be a cautionary action to ensure that some of the pages that have been added to the index before have not become spam.

There are many different professional search engine crawlers available today, such as the Google Crawler, and are used to list the URLs for use in the search engine. Without search engine crawlers, there would be no results for search engine results pages, and new pages would never be listed.

Objective –

When you hear people talk about Crawlers in the context of SEO, they are referring to the programs that search engines use to scan and analyze websites in order to determine their importance, and thus their ranking in the results of internet searches for certain keywords. Crawlers are also often referred to as spiders or robots.

Crawlers are very active, and often account for a great deal of the visitors to websites all over the internet. The Google crawler, known as Googlebot, is particularly active, and will often visit a website several times a day, checking for updates or new information. Studies have shown that it is much more active than other crawlers, the closest being the Yahoo crawler, which is about half as active as Googlebot.

Search engine optimization is aimed at understanding how crawlers work and what they look for the determine the importance and ranking of certain sites. The idea is to then implement SEO marketing strategies that will fill websites with the kind of information that the crawlers will determine to be of high value.

Crawlers are on the lookout for sites that are rich with the kinds of keywords that people search for, and sites that contain those keywords in high density are seen as being more relevant, and thus will be awarded high rankings. However, crawlers also gather other information important in determining rankings, including link population information and filenames structure.

Some forms of search engine marketing are deliberately aimed at deceiving the crawlers into thinking a site is more important than it is. These are known as black hat techniques, and they are frowned upon by most web optimizers, as they can produce penalties from search engines. There are all kinds of SEO tools out there to help you better understand crawlers and how they work. The Google keyword tool is a good place to start.

Example of Spiders –

Google began as an academic search engine. In the paper that describes how the system was built, Sergey Brin and Lawrence Page give an example of how quickly their spiders can work. They built their initial system to use multiple spiders, usually three at one time. Each spider could keep about 300 connections to Web pages open at a time.

At its peak performance, using four spiders, their system could crawl over 100 pages per second, generating around 600 kilobytes of data each second.

Modules of Web Crawler in Search Engine Optimization –

SEO is a very big, deep sea. To fully understand SEO, we should know some basic SEO terms.

**Crawling** and **indexing** are two such terms.

If you have been in the web world for a while, you’ve **at least heard** the words:

* ***Google Crawling and Indexing.***

These are the two terms upon which the **entire web world depends.**

Let’s define, understand, and get some in-depth information about crawling and indexing.

***Crawling:***

* When Google visits your website for tracking purposes. This process is done by Google’s Spider crawler.

***Indexing:***

* After crawling has been done, the results get put onto Google’s index (i.e. web search).

Crawling basically means**following a path.**

In the SEO world, crawling means**following your links** and “crawling” around your website. When bots come to your website (any page), they follow other linked pages also on your website.

This is one reason **why we create site maps,** as they contain all of the links in our blog and Google’s bots can use them to ***look deeply into a website.***

* The way we **stop crawling** certain parts of our site is by using the [Robots.txt file](https://www.shoutmeloud.com/use-robotstxt-to-protect-your-blog-from-duplicate-content-issue.html).

Factors That Affect Crawling –

There are millions of websites on this earth. Is everyone satisfied with the crawling and indexing rate?? No!! Most people are left constantly wondering why their articles aren’t getting indexed.

Let’s take a look at some major factors which play some important roles at the backend of crawling and indexing.

* 1. Domain Name
  2. Backlinks
  3. Internal Linking
  4. XML Sitemaps
  5. Duplicate Content
  6. URL Canonicalization
  7. Meta Tags
  8. Pinging

When you optimize your website based on these factors, Google will have no choice but to crawl and index your page faster and more accurately.

Architecture of Spiders or Web crawler –

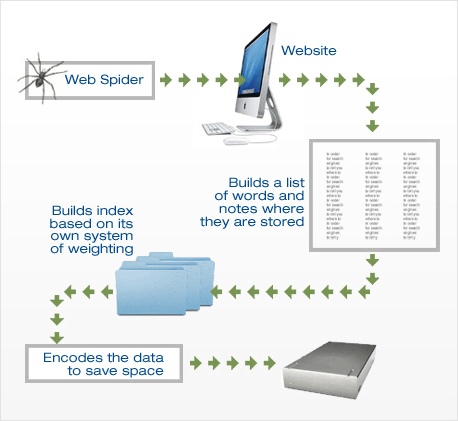


Figure 1.1 "Spiders" take a Web page's content and create key search words that enable online users to find pages they're looking for.

Conclusion –

A crawler is a program that retrieves and stores pages from the Web, commonly for a Web search engine. A crawler often has to download hundreds of millions of pages in a short period of time and has to constantly monitor and refresh the downloaded pages. In addition, the crawler should avoid putting too much pressure on the visited Web sites and the crawler's local network, because they are intrinsically shared resources.

Search engine optimization (SEO) is the process of affecting the visibility of a website or a web page in a web search engine's unpaid results—often referred to as "natural", "organic", or "earned" results. In general, the earlier (or higher ranked on the search results page), and more frequently a site appears in the search results list, the more visitors it will receive from the search engine's users; these visitors can then be converted into customers.

SEO may target different kinds of search, including image search, video search, academic search, news search, and industry-specific vertical search engines. SEO differs from local search engine optimization in that the latter is focused on optimizing a business' online presence so that its web pages will be displayed by search engines when a user enters a local search for its products or services.

System Requirements –

1. Hardware Requirements –

* Processor: Intel Dual Core or Above.
* Processor Speed: 1.0 GHz or Above.
* RAM: 256 MB or Above.
* Hard Disk: 20 GB or Above.

1. Software Requirements –

* Language : Python, Concept of HTML
* IDLE: Pycharm or Python Integrated Development Environment.
* Internet Connection for BOTs to Crawl.

Bibliography –

* How Stuff Works ( <http://computer.howstuffworks.com/internet/basics/search-engine1.htm>).
* Youtube ( <https://www.youtube.com/watch?v=CDXOcvUNBaA> ).
* Wikipedia (<https://en.wikipedia.org/wiki/Search_engine_optimization>).
* Blog-ShoutMeLoud (<https://www.shoutmeloud.com/google-crawling-and-indexing.html>)
* Search Engine land –( [www.searchengineland.com/guide/what-is-seo](http://www.searchengineland.com/guide/what-is-seo) ).