Web crawler

1.Different kind

Search engine

Copyright violation detection

Keyword based finding

Web malware detection

Web analytics

Data science data

2. feature we want to support in crawler

Politeness/ crawl rate

DNS query

Distributed crawling

Priority crawling

Duplicate detection

**Search indexing**

Search indexing is like creating a library card catalog for the Internet so that a search engine knows where on the Internet to retrieve information when a person search for it. It can also be compared to the index in the back of a book, which lists all the places in the book where a certain topic or phrase is mentioned.

Indexing focuses mostly on the text that appears on the page, and on the metadata\* about the page that users don't see. When most search engines index a page, they add all the words on the page to the index – except for words like "a," "an," and "the" in Google's case. When users search for those words, the search engine goes through its index of all the pages where those words appear and selects the most relevant ones.

\*In the context of search indexing, metadata is data that tells search engines what a webpage is about. Often the meta title and meta description are what will appear on search engine results pages, as opposed to content from the webpage that's visible to users.

**How web crawler work?**

The Internet is constantly changing and expanding. Because it is not possible to know how many total webpages there are on the Internet, web crawler bots start from a seed, or a list of known URLs. 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.

**The relative importance of each webpage**: Most web crawlers don't crawl the entire publicly available Internet and aren't intended to; instead they decide which pages to crawl first based on the number of other pages that link to that page, the amount of visitors that page gets, and other factors that signify the page's likelihood of containing important information.

**Revisiting webpages**: Content on the Web is continually being updated, removed, or moved to new locations. Web crawlers will periodically need to revisit pages to make sure the latest version of the content is indexed.

**Robots.txt requirements**: Web crawlers also decide which pages to crawl based on the robots.txt protocol (also known as the robots exclusion protocol). Before crawling a webpage, they will check the robots.txt file hosted by that page's web server. A robots.txt file is a text file that specifies the rules for any bots accessing the hosted website or application. These rules define which pages the bots can crawl, and which links they can follow.

**Bot species**:

**Search engine bots**: Also known as web crawlers or spiders: These bots "crawl," or review, content on almost every website on the Internet, and then index that content so that it can show up in search engine results for relevant user searches. They're operated by search engines like Google, Bing, or Yandex.

**Copyright bots**: Bots that crawl platforms or websites looking for content that may violate copyright law. These bots can be operated by any person or company who owns copyrighted material. Copyright bots can look for duplicated text, music, images, or even videos.

**Site monitoring bots**: These bots monitor website metrics – for example, monitoring for backlinks or system outages – and can alert users of major changes or downtime. For instance, Cloudflare operates a crawler bot called Always Online that tells the Cloudflare network to serve a cached version of a webpage if the origin server is down.

**Commercial bots**: Bots operated by commercial companies that crawl the Internet for information. These bots may be operated by market research companies monitoring news reports or customer reviews, ad networks optimizing the places where they display ads, or SEO agencies that crawl clients' websites.

**Feed bots**: These bots crawl the Internet looking for newsworthy content to add to a platform's news feed. Content aggregator sites or social media networks may operate these bots.

**Chatbots**: Chatbots imitate human conversation by answering users with preprogrammed responses. Some chatbots are complex enough to carry on lengthy conversations.

**Personal assistant bots**: like Siri or Alexa: Although these programs are much more advanced than the typical bot, they are bots nonetheless: computer programs that browse the web for data.