Search Engines

A web search engine is a software system that is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as search engine results pages (SERPs). The information may be a mix of web pages, images, and other types of files. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain real-time information by running an algorithm on a web crawler. Internet content that is not capable of being searched by a web search engine is generally described as the deep web.

# How a Search Engine Works

A search engine maintains the following processes in near real time:

* Web crawling
* Indexing
* Searching

Web search engines get their information by web crawling from site to site. The "spider" checks for the standard filename robots.txt, addressed to it, before sending certain information back to be indexed depending on many factors, such as the titles, page content, JavaScript, Cascading Style Sheets (CSS), headings, as evidenced by the standard HTML markup of the informational content, or its metadata in HTML meta tags. No web crawler may actually crawl the entire reachable web. Due to infinite websites, spider traps, spam, and other exigencies of the real web, crawlers instead apply a crawl policy to determine when the crawling of a site should be deemed sufficient. Some sites are crawled exhaustively, while others are crawled only partially.

Indexing is when the data from a crawl is processed and placed in a database. Indexing means associating words and other definable tokens found on web pages to their domain names and HTML-based fields. The associations are made in a public database, made available for web search queries. A query from a user can be a single word. The index helps find information relating to the query as quickly as possible. Some of the techniques for indexing, and caching are trade secrets, whereas web crawling is a straightforward process of visiting all sites on a systematic basis.

Between visits by the spider, the cached version of page (some or all the content needed to render it) stored in the search engine working memory is quickly sent to an inquirer. If a visit is overdue, the search engine can just act as a web proxy instead. In this case the page may differ from the search terms indexed. The cached page holds the appearance of the version whose words were indexed, so a cached version of a page can be useful to the web site when the actual page has been lost, but this problem is also considered a mild form of link rot.

Typically, when a user enters a query into a search engine it is a few keywords. The index already has the names of the sites containing the keywords, and these are instantly obtained from the index. The real processing load is in generating the web pages that are the search results list: Every page in the entire list must be weighted according to information in the indexes. Then the top search result item requires the lookup, reconstruction, and markup of the snippets showing the context of the keywords matched. These are only part of the processing each search results web page requires, and further pages (next to the top) require more of this post processing.

Beyond simple keyword lookups, search engines offer their own GUI- or command-driven operators and search parameters to refine the search results. These provide the necessary controls for the user engaged in the feedback loop users create by filtering and weighting while refining the search results, given the initial pages of the first search results. For example, from 2007, Google has allowed one to filter by date by clicking "Show search tools" in the leftmost column of the initial search results page, and then selecting the desired date range. It's also possible to weight by date because each page has a modification time.

Most search engines support the use of the Boolean operators AND, OR and NOT to help end users refine the search query. Boolean operators are for literal searches that allow the user to refine and extend the terms of the search. The engine looks for the words or phrases exactly as entered. Some search engines provide an advanced feature called proximity search, which allows users to define the distance between keywords. There is also concept-based searching where the research involves using statistical analysis on pages containing the words or phrases you search for. As well, natural language queries allow the user to type a question in the same form one would ask it to a human. A site like this would be ask.com.

The usefulness of a search engine depends on the relevance of the result set it gives back. While there may be millions of web pages that include a particular word or phrase, some pages may be more relevant, popular, or authoritative than others. Most search engines employ methods to rank the results to provide the "best" results first. How a search engine decides which pages are the best matches, and what order the results should be shown in, varies widely from one engine to another.

The methods also change over time as Internet usage changes and new techniques evolve. There are two main types of search engine that have evolved: one is a system of predefined and hierarchically ordered keywords that humans have programmed extensively. The other is a system that generates an "inverted index" by analyzing texts it locates. This first form relies much more heavily on the computer itself to do the bulk of the work.

# Market share

Google is the world's most popular search engine, with a market share of 90.14 percent as of February, 2018.

## East Asia and Russia

In some East Asian countries and Russia, Google is not the most popular search engine.

In Russia, Yandex commands a market share of 61.9 percent, compared to Google's 28.3 percent. In China, Baidu is the most popular search engine. South Korea's homegrown search portal, Naver, is used for 70 percent of online searches in the country. Yahoo! Japan and Yahoo! Taiwan are the most popular avenues for internet search in Japan and Taiwan, respectively.

## Europe

Most countries' markets in Western Europe are dominated by Google, except for Czech Republic, where Seznam is a strong competitor.

# Leading Search Engines

1. Google <https://www.google.co.in/>
2. Bing <https://www.bing.com/>
3. Yahoo! (Powered by Bing) <https://www.yahoo.com/>
4. Ask.com <https://www.ask.com/>
5. DuckDuckGo <https://www.duckduckgo.com/>
6. AOL.com <https://www.aol.com/>
7. Baidu <https://www.baidu.com/>
8. Yandex <https://www.yandex.ru/>
9. Wolfram Alpha <https://www.wolframalpha.com/>
10. Qwant <https://www.qwant.com/>

# The Google Search Engine

Google's search engine is a powerful tool. Without search engines like Google, it would be practically impossible to find the information you need when you browse the Web. Like all search engines, Google uses a special algorithm to generate search results. While Google shares general facts about its algorithm, the specifics are a company secret. This helps Google remain competitive with other search engines on the Web and reduces the chance of someone finding out how to abuse the system.

Google uses automated programs called **spiders** or **crawlers**, just like most search engines. Also like other search engines, Google has a large index of **keywords** and where those words can be found. What sets Google apart is how it ranks search results, which in turn determines the order Google displays results on its search engine results page (SERP). Google uses a trademarked algorithm called **PageRank**, which assigns each Web page a relevancy score.

A Web page's PageRank depends on a few factors:

* **The frequency and location of keywords within the Web page**: If the keyword only appears once within the body of a page, it will receive a low score for that keyword.
* **How long the Web page has existed**: People create new Web pages every day, and not all of them stick around for long. Google places more value on pages with an established history.
* **The number of other Web pages that link to the page in question**: Google looks at how many web pages link to a particular site to determine its relevance.

Because Google looks at links to a Web page as a vote, it's not easy to cheat the system. The best way to make sure your Web page is high up on Google's search results is to provide great content so that people will link back to your page. The more links your page gets, the higher its PageRank score will be. If you attract the attention of sites with a high PageRank score, your score will grow faster.

## The Fundamentals of Search

Before you search, web crawlers gather information from across hundreds of billions of web pages and organize it in the Search index. The crawling process begins with a list of web addresses from past crawls and sitemaps provided by website owners. As our crawlers visit these websites, they use links on those sites to discover other pages. The software pays special attention to new sites, changes to existing sites and dead links. Computer programs determine which sites to crawl, how often and how many pages to fetch from each site.

Google offers webmaster tools to give site owners granular choices about how Google crawls their site: they can provide detailed instructions about how to process pages on their sites, can request a recrawl or can opt out of crawling altogether using a file called “robots.txt”. Google never accepts payment to crawl a site more frequently, we provide the same tools to all websites to ensure the best possible results for our users.

## Finding information by 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.

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. When a web crawler visits a page, it collects every link on the page and adds them to its list of next pages to visit. It goes to the next page in its list, collects the links on that page, and repeats.

Web crawlers also revisit past pages once in a while to see if any changes happened. This means any site that’s linked from an indexed site will eventually be crawled. Some sites are crawled more frequently, and some are crawled to greater depths, but sometimes a crawler may give up if a site’s page hierarchy is too complex.

The web is like an ever-growing library with billions of books and no central filing system. We use software known as web crawlers to discover publicly available web pages. Crawlers look at web pages and follow links on those pages, much like you would if you were browsing content on the web. They go from link to link and bring data about those web pages back to Google’s servers.

## Organizing information by Indexing

When crawlers find a webpage, our systems render the content of the page, just as a browser does. We take note of key signals – from keywords to website freshness – and we keep track of it all in the Search index. The Google Search index contains hundreds of billions of web pages and is well over 100,000,000 gigabytes in size. It’s like the index in the back of a book – with an entry for every word seen on every web page we index. When we index a web page, we add it to the entries for all of the words it contains.

Google initiated an experiment with its search engine in 2008. For the first time, Google is allowing a group of beta testers to change the ranking order of search results. In this experiment, beta testers can promote or demote search results and tailor their search experience so that it's more personally relevant. Google executives say there's no guarantee that the company will ever implement this feature into the search engine globally.

## Specialized Searches

Google's specialized searches are an extension of its normal search engine protocol. With specialized searches, you can narrow your search to specific resources. You can enter keywords into Google and search for:

* Images related to your keywords
* Maps
* News articles or footage
* Products or services you can purchase online
* Blog entries containing the keywords you've chosen
* Content in books
* Videos
* Scholarly papers

For these searches, Google has created specialized indexes that only contain relevant sources. For example, if you search for the term "Planet Earth" in the news category, the results will include only news articles that contain those keywords. The results will look very different from Google's normal SERP.