

Web search engines work by storing information about many web pages, which they retrieve from the html itself. These pages are retrieved by a [Web crawler](#) (sometimes also known as a spider) — an automated Web browser which follows every link on the site. Exclusions can be made by the use of [robots.txt](#). The contents of each page are then analyzed to determine how it should be [indexed](#) (for example, words are extracted from the titles, headings, or special fields called [meta tags](#)). Data about web pages are stored in an index database for use in later queries. A query can be a single word. The purpose of an index is to allow information to be found as quickly as possible. Some search engines, such as [Google](#), store all or part of the source page (referred to as a [cache](#)) as well as information about the web pages, whereas others, such as [AltaVista](#), store every word of every page they find. This cached page always holds the actual search text since it is the one that was actually indexed, so it can be very useful when the content of the current page has been updated and the search terms are no longer in it. This problem might be considered to be a mild form of [linkrot](#), and Google's handling of it increases [usability](#) by satisfying [user expectations](#) that the search terms will be on the returned webpage. This satisfies the [principle of least astonishment](#) since the user normally expects the search terms to be on the returned pages. Increased search relevance makes these cached pages very useful, even beyond the fact that they may contain data that may no longer be available elsewhere.

When a user enters a [query](#) into a search engine (typically by using [key words](#)), the engine examines its [index](#) and provides a listing of best-matching web pages according to its criteria, usually with a short summary containing the document's title and sometimes parts of the text. The index is built from the information stored with the data and the method by which the information is indexed. Unfortunately, there are currently no known public search engines that allow documents to be searched by date. Most search engines support the use of the [boolean operators](#) AND, OR and NOT to further specify 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 second form relies much more heavily on the computer itself to do the bulk of the work.

Most Web search engines are commercial ventures supported by [advertising](#) revenue and, as a result, some employ the practice of allowing advertisers to [pay money to have their listings](#)

[ranked](#) higher in search results. Those search engines which do not accept money for their search engine results make money by [running search related ads](#) alongside the regular search engine results. The search engines make money every time someone clicks on one of these ads.