

Search Engines

REVIEW Page

Below is the entire module on one page. When you are ready to take your post test, scroll to the bottom of the page and press Continue>>



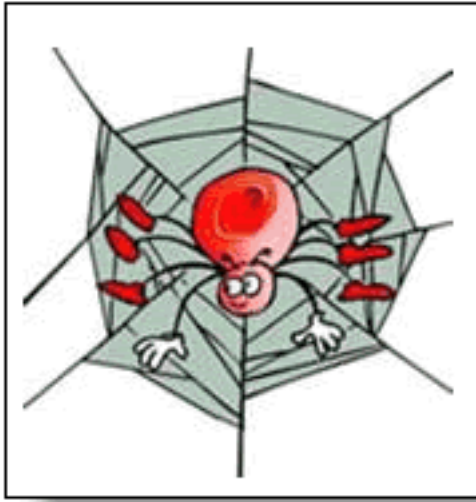
Where are you searching?

When you use a search engine you are not actually searching the live web. Instead, you are working with the search engine's database of web page information. Search engines attempt to copy and organize all of the information available on the Internet into their own efficient database. The database may include partial or complete copies of millions (billions) of web pages. Clicking on the hyperlink listed in the search results takes you from the search engine host, to the actual web page of interest.

The Search Engine three stage process: crawl, index, retrieve. Building a search engine database is a three-stage process. Search engines must find pages, organize the information for fast retrieval, and serve up the information based on your request. This is an ongoing process because the search engine databases are under constant development.

Stage 1 - Crawling:

Search engines compile huge databases of information. The database is continually updated by robotic 'crawlers or spiders' that automatically copy the contents of millions of pages of web information each day. However, not every page on every site is 'crawled and copied'. Typically a spider program takes a representative sample from a site before moving on. A spider program might copy 100 pages from a 300 page site. Valuable information may be missed. This un-indexed information is sometimes called the 'opaque' or partially visible web.



Crawlers jump from web page to web page by following each site's links. A site that is linked to many others is more likely to be visited frequently. Isolated pages will be 'crawled' less often, or can be missed altogether. This rather haphazard method of collecting information can lead to gaps in a search engine's database. This explains in part why different search engines produce different results.

It takes time and money to 'crawl' the net. Economic decisions are made that affect the depth, breadth and currency of a search engine's database. While currency and completeness are touted by each search engine, users should be aware that they are searching copies of information that could be months, even years old.

Stage 2 - Indexing:

Indexing or cataloging is the process of organizing a search engine's database of web pages to maximize retrieval efficiency. The exact methods of indexing information used by commercial search engines remain closely guarded, proprietary information. Each search engine uses a variety of 'black box' indexing algorithms to sort the contents of a page. In general the occurrence of keywords, the proximity of keywords to each other, and the contents of html elements like meta tags, titles, and headers, are all taken into account as the index is created.



The raw results delivered by the spider software, are 'pre-processed' to eliminate duplicate pages, and remove 'spam' pages. Spam pages are designed to fool search engines and achieve a higher ranking by using rigged attributes such as misleading meta-tags. Unlike the collection of human edited pages found in Subject Indexes, there is no human judgment regarding the actual quality of information. Robotic crawling and indexing is an automated process involving millions of pages of information a day.

Pages remain unavailable until they are indexed. There is lag time between the 'spidering' of a page and when it is indexed and made available to the users of the search engine. Once the web page copy is in place, it remains unchanged in the database until the next 'visit' by the search engine's crawler.

Stage 3 - Retrieving:

Finally, a search engine must provide a way for a user to retrieve information from its database. Each search engine uses a proprietary 'retrieval algorithm' to process your query. Responding to a user query is usually a two-step procedure. First the system checks for records that match your query, then it sorts or arranges the results in some kind of hierarchy. Exactly how each search engine matches queries to records is a carefully guarded trade secret.



In general search engines look for the frequency of key word matches, and the distribution of those words in a document to determine which information is relevant to your request. Key word matches found in titles and headings on the first page are given greater weight and considered semantically more relevant.

To improve retrieval speed, many search engine's filter out high frequency words such as: or, to, with, and, if, etc. These words are sometimes referred to as 'stop words'.

The most likely matches for your query are then displayed in a results list. The order in which matched pages are listed varies. Google lists results based in part on their popularity. Popularity is based on the number of other pages that link to the page in question. Commercial providers might also pay to be ranked at the top of the results list.

Would you like to see an animated overview of how search engines work? Try Learnthenet.com!

FAQ's

Why should I use more than a single Search Engine?

No single search engine covers it all. In fact there are billions of pages of information that remain hidden to search engines. Since search engines differ in their crawling, indexing, and retrieval procedures, their results will vary. While overlap exists, each search engine will contain web pages that



have been missed by the others. This means that identical queries made to different search engines will yield different results. For this reason alone, using more than a single search engine is a wise move. If you rely on a single source of information you will get an incomplete picture.

These variations occur even with search engines that use the same database of information. For example, Netscape, and AOL, use the Google Database. However each of these organizations applies their own method for retrieving information from the Google database. Additional content from their other resources may be added to the search results. This is why identical queries will often return different results.

How 'fresh' or 'current' is the information we get from a search engine?

Search engines must first find, copy, and index a web resource before it is made available for retrieval. This process takes time. Crawlers may return to a page on a daily, weekly, or monthly basis. Additionally, author submitted web pages might take weeks or months to process. Once a page is in the database, it is only a copy of the original, which may have already changed. The best way to determine the 'currency' of a web page is to examine the original material, looking for a record of when the page was last updated.

How do search engines find web pages?

We have already seen how information discovery software, sometimes-called spiders or crawlers, automates the process of finding new web pages. Most search engines also allow authors to submit their pages directly. The author supplies the web address and some information about content. The sites are then crawled, indexed and made available. Some search engines allow web page authors to buy quick placement in their systems.

How much information on each page is actually indexed?

Some search engines make a copy of the entire web page. Others take a snapshot consisting of essential address information and the first few hundred words of text on the page. There is no guarantee that all of the pages on a site have been indexed. Usually, only the main pages are included in the search

engine's database. Additionally some words are left out of the indexing process. Conjunctions, numbers and common words like 'web' or 'internet' might be excluded. These 'stop words' are removed to improve system speed and efficiency, but this practice can lead to lost information as well.

Are crawlers, spiders, and robotic information discovery the same thing? Do they work the same way?

Crawlers, spiders, and robotic information discovery all describe the process of automatic web page copying. This is an essential first step in the process of building and maintaining a search engine database. Because it is an automated process, crawlers work around the clock to find new sites and recheck sites for changes. Crawlers can be set to investigate a website in depth, visiting and copying every page. Crawlers might also just skim the surface content of a site, leaving a lot of information in the shadows.

[Authored by Dennis O'Connor 2003](#)