

Search Engines



how do they work?

Search Engines for the general web (like all those listed above) do not really search the World Wide Web directly. Each one searches a database of the full text of web pages automatically harvested from the billions of web pages out there residing on servers.

When you search the web using a search engine, you are always searching a somewhat stale copy of the real web page. When you click on links provided in a search engine's search results, you retrieve from the server the current version of the page.


Search engine databases are selected and built by computer robot programs called spiders. These "crawl" the web, finding pages for potential inclusion by following the links in the pages they already have in their database (i.e., already "know about"). They cannot think or type a URL or use judgment to "decide" to go look something up and see what's on the web about it. (Computers are getting more sophisticated all the time, but they are still brainless.)

If a web page is never linked to in any other page, search engine spiders cannot find it. The only way a brand new page - one that no other page has ever linked to - can get into a search engine is for its URL to be sent by some human to the search engine companies as a request that the new page be included. All search engine companies offer ways to do this.

After spiders find pages, they pass them on to another computer program for "indexing." This program identifies the text, links, and other content in the page and stores it in the search engine database's files so that the database can be searched by keyword and whatever more advanced approaches are offered, and the page will be found if your search matches its content.

Many web pages are excluded from most search engines by policy. The contents of most of the searchable databases mounted on the web, such as library catalogs and article databases, are excluded because search engine spiders cannot access them. All this material is referred to as the "Invisible Web" -- what you don't see in search engine results.

UC Berkeley - *Teaching Library Internet Workshops*

 is a product of

the flyinglizard

integrated marketing communication

level 4 | 9 moray place | po box 979 | dunedin | new zealand

www.goo.net.nz

Search Engines



what document features make a good match to a query?

We have discussed how search engines work, but what features of a query make for good matches? Let's look at the key features and consider some pros and cons of their utility in helping to retrieve a good representation of documents/pages.

Term frequency : How frequently a query term appears in a document is one of the most obvious ways of determining a document's relevance to a query. While most often true, several situations can undermine this premise. First, many words have multiple meanings — they are polysemous. Think of words like "pool" or "fire." Many of the non-relevant documents presented to users result from matching the right word, but with the wrong meaning. Also, in a collection of documents in a particular domain, such as education, common query terms such as "education" or "teaching" are so common and occur so frequently that an engine's ability to distinguish the relevant from the non-relevant in a collection declines sharply. Search engines that don't use a tf/idf weighting algorithm do not appropriately down-weight the overly frequent terms, nor are higher weights assigned to appropriate distinguishing (and less frequently-occurring) terms, e.g., "early-childhood."

Location of terms : Many search engines give preference to words found in the title or lead paragraph or in the metadata of a document. Some studies show that the location — in which a term occurs in a document or on a page — indicates its significance to the document. Terms occurring in the title of a document or page that match a query term are therefore frequently weighted more heavily than terms occurring in the body of the document. Similarly, query terms occurring in section headings or the first paragraph of a document may be more likely to be relevant.

Link analysis : Web-based search engines have introduced one dramatically different feature for weighting and ranking pages. Link analysis works somewhat like bibliographic citation practices, such as those used by Science Citation Index. Link analysis is based on how well-connected each page is, as defined by Hubs and Authorities, where Hub documents link to large numbers of other pages (out-links), and Authority documents are those referred to by many other pages, or have a high number of "in-links"

J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*. 1998, pp. 668-77.

Search Engines



what document features make a good match to a query?
(continued)...


Popularity : Google and several other search engines add popularity to link analysis to help determine the relevance or value of pages. Popularity utilizes data on the frequency with which a page is chosen by all users as a means of predicting relevance. While popularity is a good indicator at times, it assumes that the underlying information need remains the same.

Date of Publication: Some search engines assume that the more recent the information is, the more likely that it will be useful or relevant to the user. The engines therefore present results beginning with the most recent to the less current.

Length : While length per se does not necessarily predict relevance, it is a factor when used to compute the relative merit of similar pages. So, in a choice between two documents both containing the same query terms, the document that contains a proportionately higher occurrence of the term relative to the length of the document is assumed more likely to be relevant.

Proximity of query terms : When the terms in a query occur near to each other within a document, it is more likely that the document is relevant to the query than if the terms occur at greater distance. While some search engines do not recognize phrases per se in queries, some search engines clearly rank documents in results higher if the query terms occur adjacent to one another or in closer proximity, as compared to documents in which the terms occur at a distance.

Proper nouns sometimes have higher weights : since so many searches are performed on people, places, or things. While this may be useful, if the search engine assumes that you are searching for a name instead of the same word as a normal everyday term, then the search results may be peculiarly skewed. Imagine getting information on "Madonna," the rock star, when you were looking for pictures of madonnas for an art history class.

 is a product of

the **flyinglizard**

integrated marketing communication

level 4 | 9 moray place | po box 979 | dunedin | new zealand

www.goo.net.nz

Search Engines




Summary

The above explanation lays out the range of processing that might occur in a search engine, along with the many options that a search engine provider decides on.

The range of options may help clarify users' frequent surprise at the results their queries return. Up till now, search engine providers have mainly opted for less, versus more, complex processing of documents and queries. The typical search results therefore leave a lot of work to be done by the searcher, who must wend their way through the results, clicking on and exploring a number of documents before finding exactly what they seek. The typical evolution of products and services suggests that this status-quo will not continue. Search engines that go further in the complexity and quality of the processing performed will be rewarded with greater allegiance by searchers, as well as financially rewarding opportunities to serve as the search engine on more organizations' intranets.

"What Document Features Make a Good Match to a Query"

by Elizabeth Liddy, Director of the Center for Natural Language Processing
Professor, School of Information Studies, Syracuse University

 is a product of

the flyinglizard

integrated marketing communication

level 4 | 9 moray place | po box 979 | dunedin | new zealand

t +64 3 471 8481