

Search Engines

WEB SEARCH ENGINE

This article is about Web search engines. For search engines in general, see Search engine (computing).

A Web search engine is a search engine designed to search for information on the World Wide Web. Information may consist of web pages, images and other types of files.

Some search engines also mine data available in newsgroups, databases, or open directories. Unlike Web directories, which are maintained by human editors, search engines operate algorithmically or are a mixture of algorithmic and human input.

HISTORY

The very first tool used for searching on the Internet was Archie. The name stands for "archive" without the "v". It was created in 1990 by Alan Emtage, a student at McGill University in Montreal. The program downloaded the directory listings of all the files located on public anonymous FTP (File Transfer Protocol) sites, creating a searchable database of file names; however, Archie did not index the contents of these sites.

The rise of Gopher (created in 1991 by Mark McCahill at the University of Minnesota) led to two new search programs, Veronica and Jughead. Like Archie, they searched the file names and titles stored in Gopher index systems. Veronica (Very Easy Rodent-Oriented Net-wide Index to Computerized Archives) provided a keyword search of most Gopher menu titles in the entire Gopher listings. Jughead (Jonzy's Universal Gopher Hierarchy Excavation And Display) was a tool for obtaining menu information from specific Gopher servers. While the name of the search engine "Archie" was not a reference to the Archie comic book series, "Veronica" and "Jughead" are characters in the series, thus referencing their predecessor.

The first Web search engine was Wandex, a now-defunct index collected by the World Wide Web Wanderer, a web crawler developed by Matthew Gray at MIT in 1993. Another very early search engine, Aliweb, also appeared in 1993, and still runs today. JumpStation (released in early 1994) used a crawler to find web pages for searching, but search was limited to the title of web pages only. One of the first "full text" crawler-based search engines was WebCrawler, which came out in 1994. Unlike its predecessors, it let users search for any word in any webpage, which became the standard for all major search engines since. It was also the first one to be widely known by the public. Also in 1994 Lycos (which started at Carnegie Mellon University) was launched, and became a major commercial endeavor.

Soon after, many search engines appeared and vied for popularity. These included Magellan, Excite, Infoseek, Inktomi, Northern Light, and AltaVista. Yahoo! was among the most popular ways for people to find web pages of interest, but its search function operated on its web directory, rather than full-text copies of web pages. Information seekers could also browse the directory instead of doing a keyword-based search.

In 1996, Netscape was looking to give a single search engine an exclusive deal to be their featured search engine. There was so much interest that instead a deal was struck with Netscape by 5 of the major search engines, where for \$5Million per year each search engine would be in a rotation on the Netscape search engine page. These five engines were: Yahoo!, Magellan, Lycos, Infoseek and Excite.

Search engines were also known as some of the brightest stars in the Internet investing frenzy that occurred in the late 1990s. Several companies entered the market spectacularly, receiving record gains during their initial public offerings. Some have taken down their public search engine, and are marketing enterprise-only editions, such as Northern Light. Many search engine companies were caught up in the dot-com bubble, a speculation-driven market boom that peaked in 1999 and ended in 2001.

Search Engines

Around 2000, the Google search engine rose to prominence. [Citation needed] The company achieved better results for many searches with an innovation called PageRank. This iterative algorithm ranks web pages based on the number and Page Rank of other web sites and pages that link there, on the premise that good or desirable pages are linked to more than others. Google also maintained a minimalist interface to its search engine. In contrast, many of its competitors embedded a search engine in a web portal.

By 2000, Yahoo was providing search services based on Inktomi's search engine. Yahoo! acquired Inktomi in 2002 and Overture (which owned AlltheWeb and AltaVista) in 2003. Yahoo! switched to Google's search engine until 2004, when it launched its own search engine based on the combined technologies of its acquisitions.

Microsoft first launched MSN Search (since re-branded Live Search) in the fall of 1998 using search results from Inktomi. In early 1999 the site began to display listings from Looksmart blended with results from Inktomi except for a short time in 1999 when results from AltaVista were used instead. In 2004, Microsoft began a transition to its own search technology, powered by its own web crawler (called msnbot).

As of late 2007, Google was by far the most popular Web search engine worldwide. A number of country-specific search engine companies have become prominent; for example Baidu is the most popular search engine in the People's Republic of China.

How Web search engines work

This section does not cite any references or sources. (November 2007)

Please help improve this section by adding citations to reliable sources. Unverifiable material may be challenged and removed.

A search engine operates, in the following order

- a) Web crawling
- b) Indexing
- c) Searching

Web search engines work by storing information about many web pages, which they retrieve from the WWW itself. These pages are retrieved by a Web crawler (sometimes also known as a spider) — an automated Web browser which follows every link it sees. 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. 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. Most search engines support the use of the Boolean operators AND, OR and NOT to further specify the search query. Some search engines provide an advanced feature called proximity search which allows users to define the distance between keywords.

Search Engines

The usefulness of a search engine depends on the relevance of the result set it gives back. While there may be millions of WebPages 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.

Most Web search engines are commercial ventures supported by advertising revenue and, as a result, some employ the controversial 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.

The vast smajority of search engines is run by private companies using proprietary algorithms and closed databases, though some are open source. [Citation needed]

Revenue in the web search portals industry is projected to grow in 2008 by 13.4 percent, with broadband connections expected to rise by 15.1 percent. Between 2008 and 2012, industry revenue is projected to rise by 56 percent as Internet penetration still has some way to go to reach full saturation in American households. Furthermore, broadband services are projected to account for an ever increasing share of domestic Internet users, rising to 118.7 million by 2012, with an increasing share accounted for by fiber-optic and high speed cable lines.

Geospatially-enabled Web search engines

A recent enhancement to search engine technology is the addition of geocoding and geoparsing to the processing of the ingested documents being indexed, to enable searching within a specified locality (or region). Geoparsing attempts to match any found references to locations and places to a geospatial frame of reference, such as a street address, gazetteer locations, or to an area (such as a polygonal boundary for a municipality).[citation needed] Through this geoparsing process, latitudes and longitudes are assigned to the found places, and these latitudes and longitudes are indexed for later spatial query and retrieval. This can enhance the search process tremendously by allowing a user to search for documents within a given map extent, or conversely, plot the location of documents matching a given keyword to analyze incidence and clustering, or any combination of the two. See the list of search engines for examples of companies which offer this feature.