SEARCH ENGINE

**Directories**

A **web directory** or **link directory** is a directory on the World Wide Web. It specializes in linking to other web sites and categorizing those links.

A web directory is not a search engine and does not display lists of web pages based on keywords; instead, it lists web sites by category and subcategory. The categorization is usually based on the whole web site rather than one page or a set of keywords, and sites are often limited to inclusion in only a few categories. Web directories often allow site owners to directly submit their site for inclusion, and have editors review submissions for fitness.

RSS directories are similar to web directories, but contain collections of RSS feeds, instead of links to web sites.

Scope of listing

Most of the directories are very general in scope and list websites across a wide range of categories, regions and languages. But there are also some niche directories which focus on restricted regions, single languages, or specialist sectors. One type of niche directory with a large number of sites in existence, is the shopping Directory for example. Shopping directories specialize in the listing of retail e-commerce sites.

Examples of well known, general, web directories are Yahoo! Directory and the **Open Directory Project (ODP)**. ODP is significant due to its extensive categorization and large number of listings and its free availability for use by other directories and search engines.

However, a debate over the quality of directories and databases still continues, as search engines use ODP's content without real integration, and some experiment using clustering. There have been many attempts to make directory development easier, such as using automated submission of related links by script, or any number of available PHP portals and programs. Recently, social software techniques have spawned new efforts of categorization, with Amazon.com adding tagging to their product pages.

Directories have various features in listing, often depend upon the price paid for inclusion:

* **Free submission** – there is no charge for the review and listing of the site.
* **Reciprocal link** – a link back to the directory must be added somewhere on the submitted site in order to get listed in the directory.
* **Paid submission** – a one-time or recurring fee is charged for reviewing/listing the submitted link.
* **No Follow** – there is a rel="nofollow" attribute associated with the link, meaning search engines will give no weight to the link.
* **Featured listing** – the link is given a premium position in a category (or multiple categories) or other sections of the directory, such as the homepage.
* **Bid for position** – where sites are ordered based on bids.
* **Affiliate Links** – where the directory earns commission for referred customers from the listed websites.

Bid for Position directories

**Bid for Position directories** or also known as bidding web directories, are paid-for-inclusion web directories where the listings of websites in the directory are ordered according to their bid amount. They are special in that the more a person pays, the higher up the list of websites in the directory they go. With the higher listing, the website becomes more visible and increases the chances that visitors who browse the directory will click on the listing. There are PHP scripts (free and paid versions) for the management of bid for position directories include **phpLinkBid** (paid), Link Bid Script (free) and a modified link bid version for **phpLD** (phpLinkDirectory).

**Search Engine**

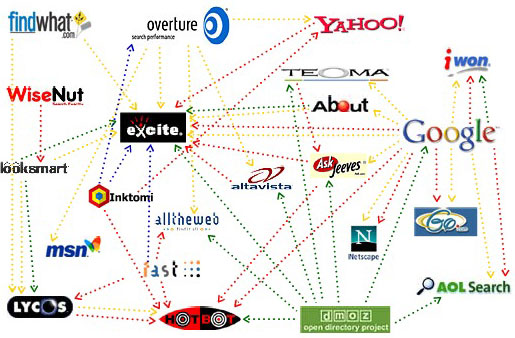
A **Web search engine** is a tool designed to search for information on the World Wide Web. The search results are usually presented in a list and are commonly called **hits**. The information may consist of web pages, images, information and other types of files. Some search engines also mine data available in newsbooks, 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

Before there were web search engines there was a complete list of all webservers. The list was edited by Tim Berners-Lee and hosted on the CERN webserver. One historical snapshot from 1992 remains. As more and more webservers went online the central list could not keep up. On the NCSA Site new servers were announced under the title "What's New!" but no complete listing existed any more.

The very first tool used for searching on the (pre-web) 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 "Jughed" are characters in the series, thus referencing their predecessor.



In June 1993, Matthew Gray, then at MIT, produced what was probably the first web root, the Perl-based World Wide Web Wanderer, and used it to generate an index called 'Wandex'. The purpose of the Wanderer was to measure the size of the World Wide Web, which it did until late 1995. The search engine Aliweb appeared in November 1993. Aliweb did not use a webroot, but instead depended on being notified by website administrators of the existence at each site of an index file in a particular format.

**Figure 1 : Placement of Search Engines**

JumpStation (released in December 1993) used a web robot to find web pages and to build its index, and used a web form as the interface to its query program. It was thus the first WWW resource-discovery tool to combine the three essential features of a web search engine (crawling, indexing, and searching) as described below. Because of the limited resources available on the platform on which it ran, its indexing and hence searching were limited to the titles and headings found in the web pages the crawler encountered.

One of the first "full text" crawler-based search engines was Web Crawler, which came out in 1994. Unlike its predecessors, it let users search for any word in any webpage, which has become 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 Alta Vista. 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.

Around 2000, the Google Search Engine rose to prominence. The company achieved better results for many searches with an innovation called Page Rank. 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 All The Weband Alta Vista 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 Alta Vista 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 and guruji.com in India.

**Meta Search Engine**

A **meta-search engine** is a search tool that sends user requests to several other search engines and/or databases and aggregates the results into a single list or displays them according to their source. Metasearch engines enable users to enter search criteria once and access several search engines simultaneously. Metasearch engines operate on the premise that the Web is too large for any one search engine to index it all and that more comprehensive search results can be obtained by combining the results from several search engines. This also may save the user from having to use multiple search engines separately.

The term Metasearch is frequently used to classify a set of commercial search engines, see the list of search engines, but is also used to describe the paradigm of searching multiple data sources in real time. The National Information Standards Organization (NISO) uses the terms Federated Search and Metasearch interchangeably to describe this web search paradigm.

Operation

Metasearch engines create what is known as a virtual database. They do not compile a physical database or catalogue of the web. Instead, they take a user's request, pass it to several other heterogeneous databases and then compile the results in a homogeneous manner based on a specific algorithm.

No two metasearch engines are alike. Some search only the most popular search engines while others also search lesser-known engines, newsgroups, and other databases. They also differ in how the results are presented and the quantity of engines that are used. Some will list results according to search engine or database. Others return results according to relevance, often concealing which search engine returned which results. This benefits the user by eliminating duplicate hits and grouping the most relevant ones at the top of the list.

Search engines frequently have different ways they expect requests submitted. For example, some search engines allow the usage of the word "AND" while others require "+" and others require only a space to combine words. The better metasearch engines try to synthesize requests appropriately when submitting them.

Quality of Results

Results can vary between metasearch engines based on a large number of variables. Still, even the most basic metasearch engine will allow more of the web to be searched at once than any one stand-alone search engine. On the other hand, the results are said to be less relevant, since a metasearch engine can’t know the internal “alchemy” a search engine does on its result (a metasearch engine does not have any direct access to the search engine’s database).

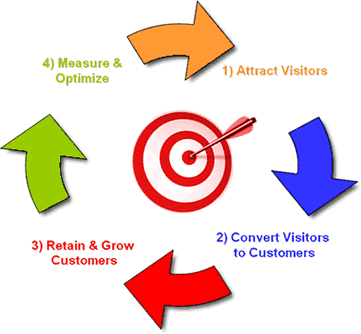
Metasearch engines are sometimes used in vertical search portals, and to search the deep web.

**Working of Search Engine**

A search engine operates, in the following order

1. **Web Crawling**
2. **Indexing**
3. **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 ofrobots.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 Alta Vista, 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.

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.

**Figure 2 : Search Engine Marketing**

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.

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.

**Submitted By :**

**Name : Rohit Aggarwal**

**Roll No. : 7CS-097**