

# History Of Search Engines

Tom Seymour, Minot State University, USA  
Dean Frantsvog, Minot State University, USA  
Satheesh Kumar, Minot State University, USA

## ABSTRACT

*As the number of sites on the Web increased in the mid-to-late 90s, search engines started appearing to help people find information quickly. Search engines developed business models to finance their services, such as pay per click programs offered by Open Text in 1996 and then Goto.com in 1998. Goto.com later changed its name to Overture in 2001, and was purchased by Yahoo! in 2003, and now offers paid search opportunities for advertisers through Yahoo! Search Marketing. Google also began to offer advertisements on search results pages in 2000 through the Google Ad Words program. By 2007, pay-per-click programs proved to be primary money-makers for search engines. In a market dominated by Google, in 2009 Yahoo! and Microsoft announced the intention to forge an alliance. The Yahoo! & Microsoft Search Alliance eventually received approval from regulators in the US and Europe in February 2010. Search engine optimization consultants expanded their offerings to help businesses learn about and use the advertising opportunities offered by search engines, and new agencies focusing primarily upon marketing and advertising through search engines emerged. The term "Search Engine Marketing" was proposed by Danny Sullivan in 2001 to cover the spectrum of activities involved in performing SEO, managing paid listings at the search engines, submitting sites to directories, and developing online marketing strategies for businesses, organizations, and individuals. Some of the latest theoretical advances include Search Engine Marketing Management (SEMM). SEMM relates to activities including SEO but focuses on return on investment (ROI) management instead of relevant traffic building (as is the case of mainstream SEO). SEMM also integrates organic SEO, trying to achieve top ranking without using paid means of achieving top in search engines, and PayPerClick SEO. For example some of the attention is placed on the web page layout design and how content and information is displayed to the website visitor.*

**Keywords:** search engines; Internet; meta-search engine; Cern

## INTRODUCTION

*W*eb Search Engine is a software program that searches the Internet (bunch of websites) based on the words that you designate as search terms (query words). Search engines look through their own databases of information in order to find what it is that you are looking for. Web Search Engines are a good example for massively sized Information Retrieval Systems.

## HISTORY

During the early development of the web, there was a list of web servers edited by Tim Berners-Lee and hosted on the CERN web server. As more web servers went online the central list could not keep up. On the NCSA site new servers were announced under the title "What's New!".

The very first tool used for searching on the Internet was Archie.<sup>1</sup>The name stands for "archive" without the "v". It was created in 1990 by Alan Emtage, Bill Heelan and J. Peter Deutsch, computer science students 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 since the amount of data was so limited it could be readily searched manually.

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.

In the summer of 1993, no search engine existed yet for the web, though numerous specialized catalogues were maintained by hand. Oscar Nierstrasz at the University of Geneva wrote a series of Perl scripts that would periodically mirror these pages and rewrite them into a standard format which formed the basis for W3Catalog, the web's first primitive search engine, released on September 2, 1993.

In June 1993, Matthew Gray, then at MIT, produced what was probably the first web robot, 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 web's second search engine Aliweb appeared in November 1993. Aliweb did not use a web robot, but instead depended on being notified by website administrators of the existence at each site of an index file in a particular format.

Jump Station (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 WebCrawler, 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 (search engine), 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 the featured search engine on Netscape's web browser. There was so much interest that instead a deal was struck with Netscape by five of the major search engines, where for \$5Million per year each search engine would be in a rotation on the Netscape search engine page. The 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, Google's search engine rose to prominence. The company achieved better results for many searches with an innovation called PageRank. This iterative algorithm ranks web pages based on the number and PageRank 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 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). Microsoft's rebranded search engine, Bing, was launched on June 1, 2009. On July 29, 2009, Yahoo! and Microsoft finalized a deal in which Yahoo! Search would be powered by Microsoft Bing technology.

## **TYPES OF SEARCH ENGINES**

### **Archie - (1990)**

History of Search Engine can be said as started in A.D. 1990. The very first tool used for searching on the Internet was Archie. It was created in 1990 by Alan Emtage, a student at McGill University in Montreal. The Archie Database was made up of the file directories from hundreds of systems. When you searched this Archie Database on the basis of a file's name, Archie could tell you which directory paths on which systems hold a copy of the file you want. Archie did not index the contents of these sites. This Archie Software, periodically reached out to all known openly available ftp sites, list their files, and build a searchable index. The commands to search Archie were UNIX commands, and it took some knowledge of UNIX to use it to its full capability.

### **Gopher - (1991)**

Later in A.D. 1991 Gopher came into the scene. Gopher was a menu system that simplified locating and using Internet resources. Gopher was designed for distributing, searching, and retrieving documents over the Internet. Gopher offered some features not natively supported by the Web and imposes a much stronger hierarchy on information stored on it. Gopher Software made it possible for the system administrator at any Internet site to prepare a customized menu of files, features and Internet resources. When you used the Gopher, all you had to do is select the item you want from the menu. Gopher was a protocol system, which in advance of the World Wide Web, allowed server based text files to be hierarchically organized and easily viewed by end users who accessed the server using Gopher Applications on remote computers. Initially Gopher Browsers could only display text-based files before developments such as Hyper Gopher, which were able to handle simple graphic formats.

### **Veronica and Jughead - (1991)**

Archie, Gopher, Veronica and Jughead were three standard "finding" tools on the Internet. The rise of Gopher led to two new search programs, Veronica and Jughead. Like Archie, they searched the file names and titles stored in Gopher index systems. Veronica was a Resource-Discovery system providing access to information resources held on most (99%+) of the world's Gopher Servers. The Veronica Database was a collection of menus from most Gopher sites. When you did a Veronica Search, you were searching the menu items. Veronica used to build an on-the-spot menu consisting of just those items that matched your request. When the search was finished, Veronica would present you with a customized Gopher menu. Veronica would not only present you with a list of Gopher menu items, it would also act like a Gopher. Jughead on the other hand was distinct from Veronica. Jughead searched a single server at a time. Jughead indexed the servers quickly so it used to build its database in memory. When Jughead used all of the available memory, it used to become unacceptably slow, limiting the size the servers it can index. Veronica does not have this problem.

### **W3Catalog & Wanderer - (1993)**

Initially, the only widely available browsers were purely textual. Mosaic was the first browser to display images in line with text instead of displaying images in a separate window. While often described as the first graphical web browser. W3Catalog was one of the first search engines that attempted to provide a general searchable catalog for WWW resources.

Unlike later search engines, like Aliweb, which attempt to index the web by crawling over the accessible content of web sites, W3Catalog exploited the fact that many high-quality, manually maintained lists of web resources were already available. W3 Catalog simply mirrored these pages, reformatted the contents into individual entries, and provided a Perl-based front-end to enable dynamic querying.

In 1993, Matthew Gray, then at MIT, produced what was probably the first web robot, the Perl-based World Wide Web Wanderer, and used it to generate an index called “Wandex”. The World Wide Web Wanderer, also referred to as just the Wanderer, was a Perl-based web crawler that was first deployed in June 1993 to measure the size of the World Wide Web. Wanderer was developed at the Massachusetts Institute of Technology by Matthew Gray, who now works for Google. It was used to generate an index called the Wandex later in 1993. While the Wanderer was probably the first web robot, and, with its index, clearly had the potential to become a general-purpose WWW search engine.

#### **Aliweb - (1993)**

Second search engine, Aliweb appeared in November 1993. Aliweb allowed users to submit the locations of index files on their sites which enabled the search engine to include WebPages and add user-written page descriptions and keywords. Aliweb, the search engine, distinguished from its contemporaries such as AltaVista by the fact that it does not automatically index sites. If a Webmaster wanted a site to be indexed by Aliweb then he or she would have to write a special file and register it with Aliweb Server. Because of the difficulty of doing this, ALIWEB has a much smaller database than search engines such as Lycos and has suffered in popularity. Aliweb provided a tool allowing users to just keep track of the services they provide, in such a way that automatic programs could simply pick up their descriptions, and combine them into a searchable database.

#### **Jump Station - (1993)**

Jump Station (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. Jump Station was thus the first WWW resource-discovery tool to combine the three essential features of a web search engine (crawling, indexing, and searching). Because of the limited resources available on the platform on which Jump Station ran, its indexing and hence searching were limited to the titles and headings found in the web pages the crawler encountered.

Jump Station used document titles and headings to index the web pages found using a simple linear search, and did not provide any ranking of results. Jump Station had the same basic shape as Google search. Brian Pinkerton, a CSE student at the University of Washington, starts WebCrawler in his spare time. At first, WebCrawler was a desktop application, not a Web service as it is today.

#### **WebCrawler - (1994)**

Brian Pinkerton, a CSE student at the University of Washington, starts WebCrawler in his spare time. At first, WebCrawler was a desktop application, not a Web service as it is today. WebCrawler went live on the Web with a database containing pages from just over 4000 different Web sites. WebCrawler was the first Web search engine to provide full text search. It went live on April 20, 1994 and was created by Brian Pinkerton at the University of Washington. It was bought by America Online on June 1, 1995 and sold to Excite on April 1, 1997. Pinkerton built a web interface to his WebCrawler program, which was released on April 20, 1994, with a database containing documents from over 6,000 web servers. The WebCrawler was unique in that it was the first web robot that was capable of indexing every word on a web page, while other bots were storing a URL, a title and at most 100 words.

#### **MetaCrawler - (1995)**

The concept of Meta-Search Engine came into existence in which a single interface provided search result that was generated by multiple search engines rather than a single Search Engine Algorithm. Daniel Dreilinger at Colorado State University developed Search Savvy which let users searched up to 20 different search engines at one and a number of directories.

MetaCrawler improved on accuracy of Search Savvy with the addition of its own search syntax and behind the scenes, matching its syntax to that of the search engines it was probing. MetaCrawler searched through six search engines, yet while providing better results, still could not match those achieved by searching each engine individually.

#### **AltaVista - (1995)**

AltaVista was once one of the most popular search engines but its popularity waned with the rise of Google. The two key participants who created the engine were Louis Monier, who wrote the crawler, and Michael Burrows, who wrote the indexer. AltaVista was backed by the most powerful computing server available. AltaVista was the fastest search engine and could handle millions of hits a day without any degradation.

One key change that came with AltaVista was the inclusion of a natural language search. Users could type in a phrase or a question and get an intelligent response. For instance, “Where is London?” without getting a million-plus pages referring to “where” and “is.”

#### **Excite - (1995)**

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 1995, Open Text provided the search technology used by Yahoo! as part of its Web index.

In 1996, Netscape was looking to give a single search engine an exclusive deal to be their featured search engine. Resultantly five major Search Engines were Yahoo!, Magellan, Lycos, Infoseek, and Excite joined the deal. Excite is an Internet portal, and as one of the major “dotcom” “portals” of the 1990s (along with Yahoo!, Lycos and Netscape), it was once one of the most recognized brands on the Internet.

Excite first appeared at the end of 1995 and was one of a spate of launches by the new ‘crawler’ based search engines, sending out spiders to record websites and build a searchable index – others from this time were AltaVista, Lycos, WebCrawler and Infoseek. SAPO was created on September 4, 1995 at the University of Aveiro by seven members of the Computer Science Center of the University.

#### **Dogpile, Inktomi, & HotBot - (1996)**

Dogpile began operation in November 1996. The site was developed by Aaron Flin. Dogpile was a metasearch site. It searched multiple engines, filtered for duplicates and then presented the results to the user. Inktomi software was incorporated in the widely-used HotBot search engine, which displaced AltaVista as the leading web-crawler-based search engine, and which was in turn displaced by Google. The company Inktomi Corporation was initially founded based on the real-world success of the search engine they developed at the university. After the bursting of the dot-com bubble, Inktomi was acquired by Yahoo!

HotBot is one of the early Internet search engines and was launched in May 1996. It updated its search database more often than its competitors. HotBot was one of the first search engines to offer the ability to search within search results. HotBot also offered free webpage hosting, but only for a short time, and it was taken down without any notice to its users. HotBot proved itself to be one of the most powerful search engines of its day, with a spider capable of indexing 10 million pages a day. This meant HotBot not only had the most up to date list of available new sites and pages, but was capable of re-indexing all previously indexed pages to ensure they were all up to date as well.

#### **Ask Jeeves & Northern Light - (1996-1997)**

Ask Jeeves (Ask) was a search engine founded in 1996 by Garrett Gruener and David Warthen in Berkeley, California. The original idea behind AskJeeves was to allow users to get answers to questions posed in everyday,

natural language, as well as traditional keyword searching. The current [Ask.com](http://Ask.com) still supports this, with added support for math, dictionary, and conversion questions.

Northern Light was to the search engine world what Apple was to the computer world. Shortly after its launch, NorthernLight like Apple, developed a fanatical following, but held a relatively small market share compared to the likes of Lycos and AltaVista. NorthernLight, from its founding in 1996 until January 2002, operated a Web search engine for public use. During this time period it also developed an enterprise offering of private custom search engines that it built for large corporate clients and marketed under the trade name Single Point. Yandex was the largest Russian Internet search engine.

#### **Google - (1998)**

Google had its rise to success in large part due to a patented algorithm called PageRank that helps rank web pages that match a given search string. Previous keyword-based methods of ranking search results, used by many search engines would rank pages by how often the search terms occurred in the page, or how strongly associated the search terms were within each resulting page. The PageRank algorithm used by Google, instead analyses human-generated links, assuming that web pages linked from many important pages are themselves likely to be important.

Google algorithm computes a recursive score for pages, based on the weighted sum of the Page Ranks of the pages linking to them. PageRank is thought to correlate well with human concepts of importance. In addition to PageRank, Google over the years has added many other secret criteria for determining the ranking of pages on result lists, reported to be over 200 different indicators. The exact percentage of the total of web pages that Google indexes are not known, as it is very hard to actually calculate. Google not only indexes and caches web pages but also takes “snapshots” of other file types, which include PDF, Word documents, Excel spreadsheets, Flash SWF, plain text files, and so on

#### **Teoma, Vivisimo - (1999-2000)**

Teoma was an Internet search engine founded in 2000 by Professor Apostolos Gerasoulis and his colleagues at Rutgers University in New Jersey. Teoma were unique because of its link popularity algorithm. Unlike PageRank Algorithm by Google, Technology of Teoma analyzed links in context to rank a web page’s importance within its specific subject. For instance, a web page about “baseball” would rank higher if other web pages about “baseball” link to it.

Vivisimo is a privately held enterprise search software company in Pittsburgh that develops and sells software products to improve search on the web and in enterprises. Vivisimo was founded in 2000 by a trio of computer science researchers at Carnegie Mellon University. Baidu was incorporated on January 18, 2000, was a Chinese and Japanese search engine for websites, audio files, and images.

Exalead was a French search engine founded in 2000 by François Bourdoncle. Exalead provides thumbnail previews of the target pages along with the results, and allows advanced refining on the results page (language, geographic location, file type, categories) but also further data refinement, such as rich content (audio, video, RSS) and related terms, allowing users to browse the web by serendipity. Info is a Metasearch Engine which provides results from leading search engines and pay-per-click directories.

#### **Yahoo! Search - (2004)**

Yahoo! Search is a web search engine, owned by Yahoo! Inc. Originally, Yahoo! Search started as a web directory of other websites, organized in a hierarchy, as opposed to a searchable index of pages. In the late 1990s, Yahoo! evolved into a full-fledged portal with a search interface. Yahoo! Search, originally referred to as Yahoo! provided Search interface, would send queries to a searchable index of pages supplemented with its directory of sites. The results were presented to the user under the Yahoo! brand.

In 2003, Yahoo! purchased Overture Services, Inc., which owned the AlltheWeb and AltaVista search engines. Initially, even though Yahoo! owned multiple search engines, they didn't use them on the main yahoo.com website, but kept using Google's search engine for its results. Starting in 2003, Yahoo! Search became its own web crawler-based search engine, with a reinvented crawler called Yahoo! Slurp. Yahoo! Search combined the capabilities of all the search engine companies they had acquired, with its existing research, and put them into a single search engine. Sogou was a Chinese search engine which can search text, images, music, and maps. It was launched 4 August 2004.

#### **MSN Search & GoodSearch - (2005)**

MSN Search was a search engine by Microsoft that comprised a search engine, index, and web crawler. MSN Search first launched in the third quarter of 1998 and used search results from Inktomi. In early 1999, MSN Search launched a version which displayed listings from Looksmart blended with results from Inktomi except for a short time in 1999 when results from AltaVista were used instead.

Since then Microsoft upgraded MSN Search to provide its own self-built search engine results, the index of which was updated weekly and sometimes daily. The upgrade started as a beta program in November 2004, and came out of beta in Feb 2005. Image search was powered by a third party, Picsearch. The service also started providing its search results to other search engine portals in an effort to better compete in the market.

GoodSearch was a Yahoo-powered search engine that donates 50% of its revenue, about a penny per search, to listed American charities and schools designated by its users. The money donated comes from the site's advertisers. SearchMe was a visual search engine based in Mountain View, California. It organized search results as snapshots of web pages.

SearchMe was founded in March 2005 by Randy Adams and John Holland.

#### **Wikiseek, Guruji, Sproose And Blackle - (2006-2007)**

Wikiseek was a search engine that indexed Wikipedia pages and pages that were linked to from Wikipedia articles. The search engine was founded by Palo Alto and was officially launched on January 17, 2007. The first public beta of Windows Live Search was unveiled on March 8, 2006, with the final release on September 11, 2006 replacing MSN Search. The new search engine used search tabs that include Web, news, images, music, desktop, local, and Microsoft Encarta.

Guruji is an Indian Internet search engine that is focused on providing better search results to Indian consumers, by leveraging proprietary algorithms and data in the Indian context. Wikia was a free and open-source Web search engine launched as part of Wikia (originally Wikicities) operated by Wikia, Inc.

Sproose is a consumer search engine launched in August 2007 by founder Bob Pack. Sproose provides web search results from partners including MSN, Yahoo! and Ask. Sproose intends to have better-quality results than algorithmic search engines because its users are able to influence the ranking order of the search results by voting for websites (which moves them up in the order of search results) and deleting bad or spam results. Blackle is a website powered by Google Custom Search, which aims to save energy by displaying a black background and using grayish-white font color for search results. The concept behind Blackle is that computer monitors can be made to consume less energy by displaying much darker colors.

#### **Powerset, Picollator, Viewzi - (2008)**

Powerset was a company based in San Francisco, California that is developing a natural language search engine for the Internet. Powerset is working on building a natural language search engine that can find targeted answers to user questions (as opposed to keyword based search). Picollator was an Internet search engine that performs search for web sites and multimedia by visual query (image) or text, or a combination of visual query and text. Picollator recognized objects in the image, obtains their relevance to the text and vice a versa, and searches in accordance with all information provided.

Viewzi is a search engine company based in Dallas, Texas that is developing a highly visual experience that tailors the way users look at information based on what they are looking for. The Viewzi search engine lightens the data overload by filtering and grouping results into several distinct interfaces.

**Cuil, LeapFish, Forestle, Valdo - (2008)**

Cuil is a search engine that organizes web pages by content and displays relatively long entries along with thumbnail pictures for many results. Boogami is a search engine that was developed by James Wildish, a sixteen year old college student from Kent in United Kingdom. It combines a search engine with a pixel advertising grid that appears every time someone uses Boogami to search the Internet, and for the fact that it offers free pixel advertising to charities.

LeapFish is a search aggregator that retrieves results from other portals and search engines, including Google, Yahoo, Live Search, Blogs, and Videos etc. Forestle is one of the ecologically inspired web search sites created by Christian Kroll, Wittenberg, Germany, in 2008. Forestle saves 0.1 square meters (about 0.1 square yards) of rain forest per search event. Valdo caters to life sciences and biomedical researchers, educators, students, clinicians and reference librarians. In addition to providing focused search on biology research methods, databases, online tools and software, Valdo is also a resource for power points on biomedical topics.

**Bing - (2009)**

Bing (formerly Live Search, Windows Live Search, and MSN Search) is a web search engine (advertised as a "decision engine") from Microsoft. Bing was unveiled by Microsoft CEO Steve Ballmer on May 28, 2009 at the All Things Digital conference in San Diego. It went fully online on June 3, 2009, with a preview version released on June 1, 2009. Notable changes include the listing of search suggestions as queries are entered and a list of related searches (called "Explorer pane") based on semantic technology from Powerset that Microsoft purchased in 2008.

**Sperse, Yebol, Goby - (2009-2010)**

Sperse Search is a metasearch engine that searches multiple search engines like Yahoo!, Bing, along with its own database and displays aggregated results, while removing duplicate results for better accuracy. Yebol is a vertical "decision" search engine that has developed a knowledge-based, semantic search platform. Based in San Jose, CA, Yebol's artificial intelligence human intelligence-infused algorithms automatically cluster and categorize search results, web sites, pages and contents that it presents in a visually indexed format that is more aligned with initial human intent. Goby is a deep web search engine which launched in September 2009. The site searches selected databases and other sources of information on the web focused on 400 categories of things to do while traveling.

**Exalead - (2011)**

Exalead is a software company that provides search platforms and search-based applications (SBA) for consumer and business users. The company is headquartered in Paris, France, and is a subsidiary of Dassault Systems. The Cloud View product is also the platform for the public Web search engine, which was designed to apply semantic processing and faceted navigation to Web data volumes and usage. Exalead also operates an online R&D laboratory, Exalabs, which uses the Web as a medium for developing applied technologies for business.

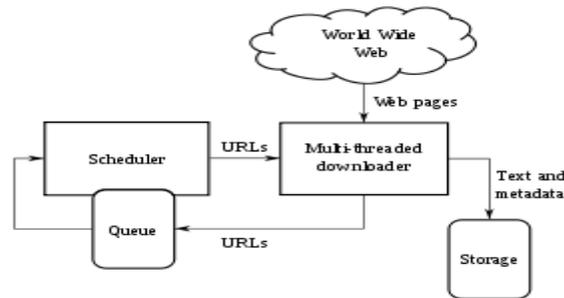
**WORKING – WEB SEARCH ENGINES**

A search engine operates, in the following order

- **Web crawling**
- **Indexing**
- **Searching**

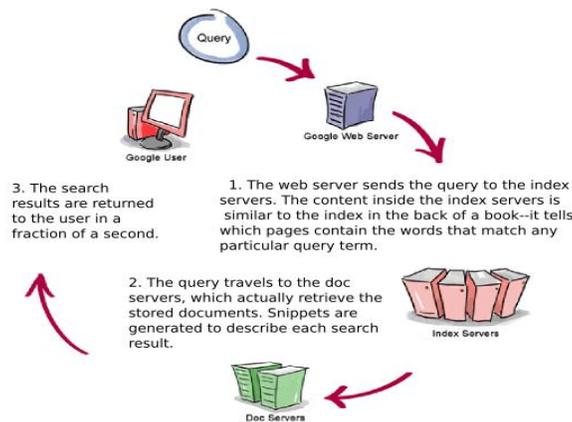
Web search engines work by storing information about many web pages, which they retrieve from the html itself. These pages are retrieved by a Web crawler (sometimes also known as a spider) — an automated Web browser which follows every link on the site. Exclusions can be made by the use of robots.txt. 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 metatags). 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.

**HIGH-LEVEL ARCHITECTURE OF STANDARD WEB CRAWLER**



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.

## **METHODS OF TEXT SEARCHING**

### **Keyword Searching**

Most search engines do their text query and retrieval using keywords. Search engines have trouble with so-called stemming.

### **Concept Searching (Clustering)**

Concept-based search systems try to determine what you mean, not just what you say. Excite is currently the best-known general-purpose search engine site on the Web that relies on concept-based searching

### **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. In to a single list or displays them according to their source. Meta-search engines do not own a database of Web pages e.g.; DOGPILE

## **MARKET SHARE & WARS**

The three most widely used web search engines and their approximate share as of late 2010. In December 2010, rankings the market share of web search engine showed Google is 91%, Yahoo is 4%, Bing is 3% and other is 2%. The Google's worldwide market share peaked at 86.3% in April, 2010. In the United States, Google held a 63.2% market share in May 2009, according to Nielsen NetRatings.<sup>1</sup> In the People's Republic of China, Baidu held a 61.6% market share for web search in July 2009.<sup>1</sup>

## **SEARCH ENGINE BIAS**

Although search engines are programmed to rank websites based on their popularity and relevancy, empirical studies indicate various political, economic, and social biases in the information they provide. These

biases could be a direct result of economic and commercial processes (e.g., companies that advertise with a search engine can become also more popular in its organic search results), and political processes (e.g., the removal of search results in order to comply with local laws). Google Bombing is one example of an attempt to manipulate search results for political, social or commercial reasons.

## CONCLUSION

The bottom line is though there are many search engines available on the web, the searching methods and the engines need to go a long way for efficient retrieval of information on relevant topics. None of the search engines out there today are perfect, but using the right one at the right time can make all the difference. Search engines need valuable websites to display in their organic search results so they can earn money from paid searches. If we build and optimize sites with visitors in mind, one can't go wrong. More often than not, what's good for website visitors is also good for the search engines. But always keep in mind, the search engines are machines without any bias toward visual effects, so it's ok to build a pretty site with lots of pictures and flash, but make sure to let the search engines know what web pages are about with an equal amount of text.

## ABOUT THE AUTHORS

**Dr. Tom Seymour** – Professor, Management Information Systems – Minot State University – Minot, ND. Dr. Tom Seymour was appointed to Chair the Business Information Technology Department at Minot State University, Minot, North Dakota for the 2007-2009 year. He has been a faculty member at MSU for 26 years. Dr. Seymour graduated from Mayville (BS), UND (MA), and Colorado State University (PhD). He came to Minot State University from Murray, Kentucky after teaching in 7 states. He is a native of Cavalier, North Dakota. He has given over 150 Computer / E-Commerce presentations in 41 states and 10 foreign countries. Dr. Seymour teaches technology classes in the classroom and via the Internet. Tom is a HLC/ NCA peer reviewer and has reviewed in 19 states including Singapore, Mexico and China. His publication record includes publishing over 80 articles in refereed journals and editing many proceedings and college textbooks. For five years Tom wrote an Internet column for the Minot Daily News and he was a North Dakota State Senator (2002-2010). Email: [tom.seymour@minotstateu.edu](mailto:tom.seymour@minotstateu.edu)

**Satheesh Kumar** has completed a bachelor of mechanical engineering degree in 1999 from one of the reputed institutions in India. He is a Senior IT Manager for Syntel Inc. (Michigan based company) and placed on an assignment at FedEx IT services in Memphis, Tennessee. Most of his experience involves providing business and technical IT solutions to Fortune 100 clients. He is completing his Master's degree in Management Information Systems at Minot State University in Minot, North Dakota. Email: [satheesh\\_sankaran@yahoo.com](mailto:satheesh_sankaran@yahoo.com)

**Dean Frantsvog** – Associate Professor, Accounting and Finance – Minot State University – Minot, North Dakota. Dean earned his Juris Doctorate degree at the Hamline University in Saint Paul, Minnesota. He teaches various business law classes in the College of Business and is a requested speaker and a Minot City Alderman. He has published various business articles and has written a book on how to start a business venture. Email: [dean.frantsvog@minotstateu.edu](mailto:dean.frantsvog@minotstateu.edu)

## REFERENCES

1. Ross, Nancy; Wolfram, Dietmar (2000). "End user searching on the Internet: An analysis of term pair topics submitted to the Excite search engine". *Journal of the American Society for Information Science* 51 (10): 949–958. doi:10.1002/1097-4571(2000)51:10<949::AID-ASI70>3.0.CO;2-5.
2. Xie, M.; et al. (1998). Quality dimensions of Internet search engines. *Journal of Information Science* 24 (5): 365–372. doi:10.1177/016555159802400509.
3. *Information Retrieval: Implementing and Evaluating Search Engines*. MIT Press. 2010.
4. Search Engine Watch [http://www.searchenginewatch.com/RFC\\_1950](http://www.searchenginewatch.com/RFC_1950) (zlib)
5. Robots Exclusion Protocol: <http://info.webcrawler.com/mak/projects/robots/exclusion.htm>
6. Web Growth Summary: <http://www.mit.edu/people/mkgray/net/web-growth-summary.html>
7. Yahoo! <http://www.yahoo.com/>

8. [Abiteboul 97] Serge Abiteboul and Victor Vianu, Queries and Computation on the Web. Proceedings of the International Conference on Database Theory. Delphi, Greece 1997.
9. [Bagdikian 97] Ben H. Bagdikian. The Media Monopoly. 5th Edition. Publisher: Beacon, ISBN: 0807061557 [Cho 98] Junghoo Cho, Hector Garcia-Molina, And Lawrence Page. Efficient Crawling Through URL Ordering. Seventh International Web Conference (WWW 98). Brisbane, Australia, April 14-18, 1998.
10. [Gravano 94] Luis Gravano, Hector Garcia-Molina, and A. Tomasic. The Effectiveness of GLOSS for the Text-Database Discovery Problem. Proc. of the 1994 ACM SIGMOD International Conference on Management of Data, 1994.
11. [Kleinberg 98] Jon Kleinberg, Authoritative Sources in a Hyperlinked Environment, Proc. ACM-SIAM Symposium on Discrete Algorithms, 1998.
12. [Marchiori 97] Massimo Marchiori. The Quest for Correct Information on the Web: Hyper Search Engines. The Sixth International WWW Conference (WWW 97). Santa Clara, USA, April 7-11, 1997.
13. [McCray 94] Oliver A. McCray. GENVL and WWW: Tools for Taming the Web. First International Conference on the World Wide Web. CERN, Geneva (Switzerland), May 25-26 1994.  
<http://www.cs.colorado.edu/home/mcbryan/mypapers/www94.ps>
14. [Page 98] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Manuscript in progress.  
<http://google.stanford.edu/~backrub/pageranksub.ps>
15. [Pinkerton 94] Brian Pinkerton, Finding What People Want: Experiences with the WebCrawler. The Second International WWW Conference Chicago, USA, October 17-20, 1994.  
<http://info.webcrawler.com/bp/WWW94.html>
16. [Spertus 97] Ellen Spertus. ParaSite: Mining Structural Information on the Web. The Sixth International WWW Conference (WWW 97). Santa Clara, USA, April 7-11, 1997.
17. [TREC 96] Proceedings of the fifth Text Retrieval Conference (TREC-5). Gaithersburg, Maryland, November 20-22, 1996. Publisher: Department of Commerce, National Institute of Standards and Technology. Editors: D. K. Harman and E. M. Voorhees. Full text at: <http://trec.nist.gov/>
18. [Witten 94] Ian H Witten, Alistair Moffat, and Timothy C. Bell. Managing Gigabytes: Compressing and Indexing Documents and Images. New York: Van Nostrand Reinhold, 1994.
19. [Weiss 96] Ron Weiss, Bienvenido Velez, Mark A. Sheldon, Chanathip Manprempre, Peter Szilagyi, Andrzej Duda, and David K. Gifford. HyPursuit: A Hierarchical Network Search Engine That Exploits Content-Link Hypertext Clustering. Proceedings of the 7th ACM Conference on Hypertext. New York, 1996.
20. Bing Liu (2007), *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data*. Springer, ISBN 3540378812
21. Bar-Ilan, J. (2004). The use of Web search engines in information science research. ARIST, 38, 231-288.
22. Levene, Mark (2005). *An Introduction to Search Engines and Web Navigation*. Pearson.
23. Hock, Randolph (2007). *The Extreme Searcher's Handbook*. ISBN 978-0-910965-76-7
24. Javed Mostafa (February 2005). "Seeking Better Web Searches". *Scientific American Magazine*.  
<http://www.sciam.com/article.cfm?articleID=0006304A-37F4-11E8-B7F483414B7F0000>