History of Search Engines: From 1945 to Google Today

As We May Think (1945):

The concept of hypertext and a memory extension really came to life in July of 1945, when after enjoying the scientific camaraderie that was a side effect of WWII, Vannevar Bush's *As We May Think* was published in The Atlantic Monthly.

He urged scientists to work together to help build a body of knowledge for all mankind. Here are a few selected sentences and paragraphs that drive his point home.

> Specialization becomes increasingly necessary for progress, and the effort to bridge between disciplines is correspondingly superficial.
>
> The difficulty seems to be, not so much that we publish unduly in view of the extent and variety of present day interests, but rather that **publication has been extended far beyond our present ability to make real use of the record**. The summation of human experience is being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square-rigged ships.
>
> A record, if it is to be useful to science, must be continuously extended, it must be stored, and above all it must be consulted.

He not only was a firm believer in storing data, but he also believed that if the data source was to be useful to the human mind we should have it represent how the mind works to the best of our abilities.

> **Our ineptitude in getting at the record is largely caused by the artificiality of the systems of indexing**. ... Having found one item, moreover, one has to emerge from the system and re-enter on a new path.
>
> **The human mind does not work this way. It operates by association.** ... Man cannot hope fully to duplicate this mental process artificially, but he certainly ought to be able to learn from it. In minor ways he may even improve, for his records have relative permanency.
>
> Presumably man's spirit should be elevated if he can better review his own shady past and analyze more completely and objectively his present problems. He has built a civilization so complex that he needs to mechanize his records more fully if he is to push his experiment to its logical conclusion and not merely become bogged down part way there by overtaxing his limited memory.

He then proposed the idea of a virtually limitless, fast, reliable, extensible, associative

Gerard Salton (1960s - 1990s):

Gerard Salton, who died on August 28th of 1995, was the father of modern search technology. His teams at Harvard and Cornell developed the SMART informational retrieval system. Salton's Magic Automatic Retriever of Text included important concepts like the vector space model, Inverse Document Frequency (IDF), Term Frequency (TF), term discrimination values, and relevancy feedback mechanisms.

He authored a 56 page book called A Theory of Indexing which does a great job explaining many of his tests upon which search is still largely based. Tom Evslin posted a blog entry about what it was like to work with Mr. Salton.

Ted Nelson:

Ted Nelson created Project Xanadu in 1960 and coined the term hypertext in 1963. His goal with Project Xanadu was to create a computer network with a simple user interface that solved many social problems like attribution.

While Ted was against complex markup code, broken links, and many other problems associated with traditional HTML on the WWW, much of the inspiration to create the WWW was drawn from Ted's work.

There is still conflict surrounding the exact reasons why Project Xanadu failed to take off.

The Wikipedia offers background and many resource links about Mr. Nelson.

Advanced Research Projects Agency Network:

ARPANet is the network which eventually led to the internet. The Wikipedia has a great background article on ARPANet and Google Video has a free interesting video about ARPANet from 1972.

Archie (1990):

The first few hundred web sites began in 1993 and most of them were at colleges, but long before most of them existed came Archie. The first search engine created was Archie, created in 1990 by Alan Emtage, a student at McGill University in Montreal. The original intent of the name was "archives," but it was shortened to Archie.

Archie helped solve this data scatter problem by combining a script-based data gatherer with a regular expression matcher for retrieving file names matching a user query. Essentially Archie became a database of web filenames which it would match with the users queries.

Bill Slawski has more background on Archie here.

Veronica & Jughead:

As word of mouth about Archie spread, it started to become word of computer and Archie had such popularity that the University of Nevada System Computing Services group developed Veronica. Veronica served the same purpose as Archie, but it worked on plain text files. Soon another user interface name Jughead appeared with the same purpose as Veronica, both of these were used for files sent via Gopher, which was created as an Archie alternative by Mark McCahill at the University of Minnesota in 1991.

File Transfer Protocol:

Tim Burners-Lee existed at this point, however there was no World Wide Web. The main

If you had a file you wanted to share you would set up an FTP server. If someone was interested in retrieving the data they could using an FTP client. This process worked effectively in small groups, but the data became as much fragmented as it was collected.

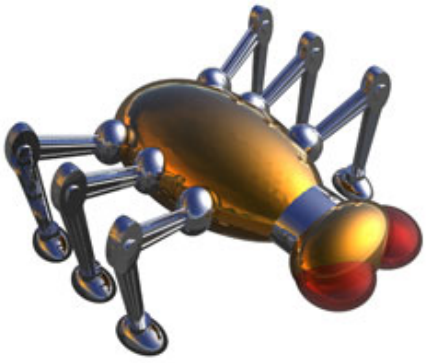
Tim Berners-Lee & the WWW (1991):

From the Wikipedia:

> While an independent contractor at CERN from June to December 1980, Berners-Lee proposed a project based on the concept of hypertext, to facilitate sharing and updating information among researchers. With help from Robert Cailliau he built a prototype system named Enquire.
>
> After leaving CERN in 1980 to work at John Poole's Image Computer Systems Ltd., he returned in 1984 as a fellow. In 1989, CERN was the largest Internet node in Europe, and Berners-Lee saw an opportunity to join hypertext with the Internet. In his words, "*I just had to take the hypertext idea and connect it to the TCP and DNS ideas and — ta-da! — the World Wide Web*". He used similar ideas to those underlying the Enquire system to create the World Wide Web, for which he designed and built the first web browser and editor (called WorldWideWeb and developed on NeXTSTEP) and the first Web server called httpd (short for HyperText Transfer Protocol daemon).
>
> The first Web site built was at http://info.cern.ch/ and was first put online on August 6, 1991. It provided an explanation about what the World Wide Web was, how one could own a browser and how to set up a Web server. It was also the world's first Web directory, since Berners-Lee maintained a list of other Web sites apart from his own.
>
> In 1994, Berners-Lee founded the World Wide Web Consortium (W3C) at the Massachusetts Institute of Technology.

Tim also created the Virtual Library, which is the oldest catalogue of the web. Tim also wrote a book about creating the web, titled *Weaving the Web*.

What is a Bot?



Computer robots are simply programs that automate repetitive tasks at speeds impossible for humans to reproduce. The term bot on the internet is usually used to describe anything that interfaces with the user or that collects data.

Search engines use "spiders" which search (or spider) the web for information. They are software programs which request pages much like regular browsers do. In addition to

- Link anchor text may help describe what a page is about.
- Link co citation data may be used to help determine what topical communities a page or website exist in.
- Additionally links are stored to help search engines discover new documents to later crawl.

Another bot example could be Chatterbots, which are resource heavy on a specific topic. These bots attempt to act like a human and communicate with humans on said topic.

Parts of a Search Engine:

Search engines consist of 3 main parts. Search engine **spiders** follow links on the web to request pages that are either not yet indexed or have been updated since they were last indexed. These pages are crawled and are added to the search engine **index** (also known as the catalog). When you search using a major search engine you are not actually searching the web, but are searching a slightly outdated index of content which roughly represents the content of the web. The third part of a search engine is the **search interface and relevancy software**. For each search query search engines typically do most or all of the following

- Accept the user inputted query, checking to match any advanced syntax and checking to see if the query is misspelled to recommend more popular or correct spelling variations.
- Check to see if the query is relevant to other vertical search databases (such as news search or product search) and place relevant links to a few items from that type of search query near the regular search results.
- Gather a list of relevant pages for the organic search results. These results are ranked based on page content, usage data, and link citation data.
- Request a list of relevant ads to place near the search results.

Searchers generally tend to click mostly on the top few search results, as noted in this article by **Jakob Nielsen**, and backed up by this search result eye tracking study.

Want to learn more about how search engines work?

- In *How does Google collect and rank results?* Google engineer Matt Cutts briefly discusses how Google works.
- Google engineer Jeff Dean lectures a University of Washington class on how a search query at Google works in this video.
- The Chicago Tribune ran a special piece titled *Gunning for Google*, including around a dozen audio interviews, 3 columns, and this graphic about how Google works.
- How Stuff Works covers search engines in *How Internet Search Engines Work*.

Types of Search Queries:

Andrei Broder authored *A Taxonomy of Web Search* [PDF], which notes that most searches fall into the following 3 categories:

- Informational - seeking static information about a topic
- Transactional - shopping at, downloading from, or otherwise interacting with the result
- Navigational - send me to a specific URL

Improve Your Searching Skills:

Want to become a better searcher? Most large scale search engines offer:

- **Advanced search pages** which help searchers refine their queries to request files which are newer or older, local or in nature, from specific domains, published in specific formats, or other ways of refining search, for example the ~ character means related to Google.
- **Vertical search databases** which may help structure the information index or limit the search index to a more trusted or better structured collection of sources, documents, and information.

Nancy Blachman's *Google Guide* offers searchers free Google search tips, and Greg R.Notess's *Search Engine Showdown* offers a search engine features chart.

There are also many popular smaller vertical search services. For example, Del.icio.us allows you to search URLs that users have bookmarked, and Technorati allows you to search blogs.

World Wide Web Wanderer:

count active web servers. He soon upgraded the bot to capture actual URL's. His database became knows as the Wandex.

The Wanderer was as much of a problem as it was a solution because it caused system lag by accessing the same page hundreds of times a day. It did not take long for him to fix this software, but people started to question the value of bots.

ALIWEB:

In October of 1993 Martijn Koster created Archie-Like Indexing of the Web, or ALIWEB in response to the Wanderer. ALIWEB crawled meta information and allowed users to submit their pages they wanted indexed with their own page description. This meant it needed no bot to collect data and was not using excessive bandwidth. The downside of ALIWEB is that many people did not know how to submit their site.

Robots Exclusion Standard:

Martijn Kojer also hosts the web robots page, which created standards for how search engines should index or not index content. This allows webmasters to block bots from their site on a whole site level or page by page basis.

By default, if information is on a public web server, and people link to it search engines generally will index it.

In 2005 Google led a crusade against blog comment spam, creating a nofollow attribute that can be applied at the individual link level. After this was pushed through Google quickly changed the scope of the purpose of the link nofollow to claim it was for any link that was sold or not under editorial control.

Primitive Web Search:

By December of 1993, three full fledged bot fed search engines had surfaced on the web: JumpStation, the World Wide Web Worm, and the Repository-Based Software Engineering (RBSE) spider. JumpStation gathered info about the title and header from Web pages and retrieved these using a simple linear search. As the web grew, JumpStation slowed to a stop. The WWW Worm indexed titles and URL's. The problem with JumpStation and the World Wide Web Worm is that they listed results in the order that they found them, and provided no discrimination. The RSBE spider did implement a ranking system.

Since early search algorithms did not do adequate link analysis or cache full page content if you did not know the exact name of what you were looking for it was extremely hard to find it.

Excite:

Excite came from the project Architext, which was started by in February, 1993 by six Stanford undergrad students. They had the idea of using statistical analysis of word relationships to make searching more efficient. They were soon funded, and in mid 1993 they released copies of their search software for use on web sites.

Excite was bought by a broadband provider named @Home in January, 1999 for $6.5 billion, and was named Excite@Home. In October, 2001 Excite@Home filed for bankruptcy. InfoSpace bought Excite from bankruptcy court for $10 million.

Web Directories:

VLib:

When Tim Berners-Lee set up the web he created the Virtual Library, which became a loose confederation of topical experts maintaining relevant topical link lists.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.