



Google - Behind the Homepage

A presentation by Betsy Chessler,
Associate Librarian, Morrison & Foerster LLP.
September 8, 2006
SLA San Diego Chapter Meeting

Overview

Google History

Developed in 1998 by Sergey Brin and Larry Page, two Stanford computer science Ph.D. students who were dissatisfied with existing search engines. The existing search engines returned too many irrelevant results. Ranking of results was also easily manipulated by Webmasters.

What is the perfect search engine? Something that “understands exactly what you mean and gives you back exactly what you want.” (- Larry Page, <http://www.google.com/corporate/tech.html>)

"We believed we could build a better search. We had a simple idea, that not all pages are created equal. Some are more important."
(- Sergey Brin http://www.berkeley.edu/news/media/releases/2005/10/04_servey.shtml)

"People are still only willing to look at the first few tens of results. Because of this, as the collection grows, we need tools that have very high precision... Indeed, we want our notion of “relevant” to only include the very best documents, since there may be tens of thousands of slightly relevant documents."
(- Sergey Brin and Lawrence Page, <http://www-db.stanford.edu/~backrub/google.html>)

How Google Works - Text analysis and PageRank

Google looks at your search terms within the context of the Web page in which they appear and considers the following when ranking search results:

A. presence of search terms - All your search terms must appear somewhere in the document, unless you specified otherwise.

B. proximity of search terms - How often do your search terms appear? How close together are they?

C. text characteristics - Are your search terms in bold text? In the header text? In a larger font size relative to the rest of the page? If so, these pages will rank higher.

D. hyperlinks at site - Google also looks at the content of neighboring Web pages to see if those pages are on the same topic. This provides another check on text analysis (see <http://www.google.com/corporate/tech.html>)

Text analysis is then combined with:

E. PageRank - A constantly tweaked algorithm that determines both the popularity and quality of any given Web page. Each page is given a rank based on the number and quality of the pages that link to it. (This may be familiar to some from "citation analysis" in the sciences and social sciences. A paper is considered more important and groundbreaking if a lot of authors cite to it.)

Just for fun, here's the algorithm as it existed in 1998:

$$PR(A) = (1-d) + d(PR(T1)/C(T1) + \dots + PR(Tn)/C(Tn))$$

<http://www-db.stanford.edu/~backrub/google.html>

Don't worry - you don't have to understand it. Check out this alternative, and more enjoyable, explanation of PageRank, or "PigeonRank":

<http://www.google.com/technology/pigeonrank.html>

Google Statistics

Bigness

As of Nov. 2004, the Google index contained 8 billion Web pages

(<http://www.google.com/corporate/history.html> and
<http://searchenginewatch.com/reports/article.php/2156481#key>).

Google recently declared it was three times larger than its closer competitor. (Sep. 2005) (<http://searchenginewatch.com/searchday/article.php/3551586>)

It is one of the 5 most popular sites on the Internet.

(see <http://www.google.com/corporate/facts.html>)

Nielsen Net Ratings tracks search behavior of 1 million+ representative users every month. In July 2006, Google was used the most among search engines (49% of the time), followed distantly by Yahoo (24%) and MSN (10%). An estimated 2.8 billion searches were conducted using Google in July 2006.

(see http://www.nielsen-netratings.com/pr/pr_060821.pdf)

Google is now so associated with Internet searching that it is listed in the dictionary. In June 2006, the Oxford English Dictionary added "Google" as a verb. In July 2006, Merriam-Webster followed suit, adding "google" with a lower case "g". (see <http://www.merriam-webster.com/dictionary/google> and
<http://www.stuff.co.nz/stuff/print/0,1478,3728360a28,00.html>)

Freshness

Why is it so important to index the Web frequently? According to the *Journal of Information Science*, an estimated 320 million new pages are added to the Web every week. About 20% of today's Web pages will disappear within the year. About 50% of the content of Web pages will change within a year. About 80% of all links will change or be new within a year's time. Luckily, Google is keeping up nicely. 83% of Google results were not more than a day old. Much better than MSN (48%) or Yahoo (42%) On average, a page is re-indexed by Google every 3 days.

(Lewandowski, et al, "The freshness of Web search engine databases", *Journal of Information Science*, 32(2) 2006, pp131-148
preprint at: http://www.durchdenken.de/lewandowski/doc/jis_preprint.pdf)

Basic Search Features

Google lets you choose between the basic search screen or the advanced search screen. You can pretty much do the exact same searches from either screen. The advantage of the advanced search screen is that you don't have to remember all your search options; simply fill in the boxes and select from drop down menus. (However, all of the examples below use the basic search screen.)

1. Google is NOT case sensitive - *NeXT* is searched the same as *Next* or *next*. Google ignores most punctuation. Hyphenated words are searched both with and without hyphen. Possessives and contractions are searched with and without the apostrophe. Example: *e-mail* *people's*
2. A space is an assumed "and". Don't type AND!
3. Common words, single digits and letters are not searched ("stop words"). Place a plus sign (+) in front of a stop word to search it or use quotes: *+Z corporation* ; *"Z Corporation"*
4. The order you type your search terms affects how your results are ranked. Compare results for: *grass snake* and *snake grass*. The top hits for the first search are about the reptile, the second about horsetail grass. (search example from Joe Barker's excellent class, Extreme Googling 2005, http://www.infopeople.org/training/past/2005/extr_googling/)
5. To force Google to search your terms exactly as entered, use quotation marks (" "). You can also use quotation marks for common words that would otherwise be ignored, e.g. *"to be or not to be"*, *"university of california"* (the latter search will avoid picking up California state universities)
6. Use OR (capitalized) to search synonyms: ex. Barrister OR attorney OR lawyer
7. The tilde (~) also searches for synonyms, example: *~attorney* also pulls up legal, lawyer, law, etc.
8. A minus sign (-) excludes words; example: *rico -puerto*
9. Google searches only first 10 words of query; anything else is ignored.
10. Google automatically searches for variations of your search terms ("stemming"). You type in "library", Google will also search for "libraries". However, your exact term ("library") will generally rank higher in search results. Example: *colloid gold*
11. Proximity searching. You can't (yet) tell Google that your search terms should appear within a couple words of each other. You can use an asterisk (*) within a phrase search as a wild card, to match any word in that position. So, for example, to find San Diego libraries when you don't know if they will be called San Diego "public" libraries or San Diego "county" libraries, search "san diego * library".

An independent programmer has developed "GAPS" (Google API Proximity Search), based on this wild card feature. It allows you to specify that your search terms appear within up to 3 words of each other. See <http://www.staggernation.com/cgi-bin/gaps.cgi>. Caution: this site is not always operational! Example: *firstname* within 2 words of *lastname* in either order (to pick up results with middle names).
12. Caching. Google saves a copy of all pages it indexes. Very useful when the actual Web page has disappeared. It only stores the page as it appeared when it was last indexed. The cached copy also highlights your search terms. For historical versions of a Web page, check out the Internet Archive (<http://www.archive.org>)

Advanced Search Features

1. Operators.

Use operators in conjunction with your search terms. Here are a few of our favorites. (Full list of operators at: <http://www.google.com/help/operators.html>)

define: Google's built-in dictionary. Will pull dictionary, glossary and encyclopedia definitions of your specified term. Example: *define:carpe diem*

filetype: Searches only file format you specify. Google can search 13 different file types in addition to .html, including .pdf, .rtf, .xls, .doc, and .ppt. To search for Powerpoint presentations on the subject of Google, type: *filetype:ppt google* See http://www.google.com/help/faq_filetypes.html for the full list of file types indexed.

link: Shows Web pages that link to your specified page. Great way to track who links to your pages.

Example: *link:mofocom*

phonebook: Looks up phone number for a business or person. Type in name of person or business and city or zip code

Examples: *phonebook:new york pizza and italian deli san diego*

phonebook:bill smith escondido

phonebook:chessler 98040

related: Shows Web pages that are similar to your specified page.

Example: *related:boeing.com*

site: Restricts your search to specified Web site or domain. Great tool if a Web site is hard to navigate or does not have its own search engine.

Examples: *site:irs.gov community property states*

site:lasuperiorcourt.org "motion to strike"

northern spotted owl site:edu

2. Calculator

Google has a built-in calculator function. Complete instructions are at: <http://www.google.com/help/calculator.html>. Great for unit and currency conversions and any arithmetic function. Examples: 12 kilometers in miles; 18 Celsius in Fahrenheit; 100 kph in mph; 5+2*3; half a cup in teaspoons; square root of 36, 1 dollar in british pounds, etc.

3. Number searches

Google offers a fast way to search specialized numbers, such as U.S. patent numbers, telephone area codes, Vehicle Identification Numbers (VINs), FedEx and UPS parcel tracking IDs, and more. See <http://www.google.com/help/features.html#number> for the full list and search examples.

Example: To search the area code 619, type: *619*

To search for U.S. patent 5123123, type: *patent 5123123*

4. Language Tools

Google will automatically translate many Web pages in foreign languages into English. Currently Google will translate Arabic, Chinese (Simplified), French, German, Italian, Korean, Japanese, Spanish, and Portuguese. The translation is done by computer

without human intervention. Results are generally satisfactory, if not always elegant. If a translation is available, you will see a link to the translation from your initial search results page. See http://www.google.com/help/faq_translation.html for more information.

Example: *der Spiegel*

To translate small portions of text to and from English, Arabic, Chinese (Simplified), French, German, Italian, Korean, Japanese, Spanish, and Portuguese, go to Google's Language Tools page (http://www.google.com/language_tools?hl=en). You can also type in a Web page and have it translated for you. Any text that appears in graphics will not be translated.

Collections

Google Images

Google has pulled billions of images from Web space and made them searchable. Click on the Images tab from the Basic search screen or go to: <http://images.google.com> Images will be presented thumbnail size for ease of viewing, but you can always enlarge the picture and see the picture as it appears within its original document. How does Google do it? "Google analyzes the text on the page adjacent to the image, the image caption and dozens of other factors to determine the image content. Google also uses sophisticated algorithms to remove duplicates and ensure that the highest quality images are presented first in your results."

http://images.google.com/help/faq_images.html

Search example: *cessna*

Google News

Search a month's worth of news articles from 4,500 sources. Updated continuously. No human editors organize the news; it is entirely computer generated. Use free Alerts service to track news on topic of your choice. Access News by clicking on the News tab at Google's main search screen or go directly to: <http://news.google.com>

Google News Archive was added in September 2006. See http://news.google.com/archivesearch/advanced_search

Search example: *time warner*

Google Maps (launched 2/8/05)

View maps and get directions. Zoom in and out. View maps in satellite or street format, or a hybrid. Search by address, city, business name, business type. Access at: <http://maps.google.com>

Search examples: *bakeries near 425 market st, san Francisco*

Hotels near seatac

Google Blog Search (launched 9/13/05)

Searches blogs (ubiquitous, self-published "Internet diaries"). Google checks blogs frequently for new content. If there is a blog that is completely devoted to your subject matter, it will be listed separately at the top of your search results.

Access at: <http://blogsearch.google.com> or <http://search.blogger.com>

Search example: *FOIA*

Google Book Search (launched 11/3/05)

Search the full text of books, then see your search words in context. If the book is copyrighted, you'll only see a few sentences, paragraph, or few pages containing your search terms. Nor can you print or cut and paste copyrighted works. If it's publicly available without restriction, you can view the entire book. Books are provided to Google for scanning by publishers and libraries. Google Book (formerly known as Google Print) is still in beta; more and more books are being added to the database.

See press release: http://www.google.com/press/pressrel/print_publicdomain.html

Access at: <http://print.google.com>

Search example: *cybersquatting*

Google Scholar (launched 3/2/06)

Search for scholarly journal articles, theses, books and abstracts. Different versions of the same article are grouped together, and you can link to a list of articles that cite to your article (very cool!). If the full text of an article is available on the Web, there will be a link to it. If known, Google will tell you which libraries have a copy of the article and/or will provide a link to British Library Direct if you want to purchase a copy of the article. You can not yet sort articles by date, but you can view newer articles by clicking on the "recent articles" hyperlink.

Access at: <http://scholar.google.com> help pages at: <http://scholar.google.com/scholar/help.html>

Search example: *colloidal gold marker*
author:"jm olefsky"

Google Finance (launched 3/21/06)

Google introduced Google finance to compete with Yahoo's popular Finance site. Google offers more interactive stock quotes, faster access to news stories affecting a company's stock price, profiles and photos of top executives, and links to blogs on the company. Google purchases its financial data from third parties, including Hoover's, Interactive Data, Reuters, and Morningstar.

Access at: <http://finance.google.com>

Search example: *goog* (ticker symbol for Google)

Google Spreadsheet (launched 6/6/06)

Still in beta mode. Free spreadsheet program that you can share and edit with up to 10 others simultaneously. Has built in chat so you can collaborate.

Access at: <http://spreadsheet.google.com>

And just announced on 9/4/06: Google "Apps for your Domain", a free suite of office applications, including Google email (Gmail), Google Calendar, Google Talk, Google Page Creator (HTML editor), and, coming soon, a word processing program. (Google acquired "Writely" in March 2006). See <https://www.google.com/a/> for more details.

What Google Can't Find and What to Do About It

A significant percentage of the Web is not indexed, either because sites are behind firewalls or require passwords or the information is contained in databases that a search engine cannot index. It's estimated that about 5% of the Internet just isn't accessible, so it can't be indexed - . See <http://searchenginewatch.com/searchday/article.php/2159121>)

Sometimes the Web page is up for such a short time that a search engine doesn't have time to index it. Sometimes a search engine just misses a site entirely.

1. To be more comprehensive, try your search in any other search engine. There is not as much overlap as you would think. (some other search engines: <http://www.yahoo.com>, <http://www.alltheweb.com>, <http://www.dogpile.com>)
2. For defunct Web pages, try the Internet Archive (<http://www.archive.org>). For defunct government agencies, also try CyberCemetery (<http://govinfo.library.unt.edu/default.htm>).
3. Search subject-specific databases, directories and search engines. (examples: Daypop, Ingenta, Highbeam Research, Hoover's, Libdex, Infomine, Pubmed, Topica, InfoRetrieve Article Finder). For finding people, try ZoomInfo (<http://www.zoominfo.com>) and PooGee (<http://www.poogee.com>).
4. Add words like "faq" "database" "directory" "expert" "pathfinder" "review" to your search terms to find specialized directories and databases on your topic. Example: health care database See: <http://marylaine.com/exlibris/xlib94.html>

Remember - much dusty old material has not and may never be digitized - you must visit a library to see and touch it. **When in doubt, ask a librarian.**

Author's note: This presentation has its genesis in a talk I did in May 2002 with Judy Broom and Katherine Foster for the Law Librarians of Puget Sound. Our original presentation outline is at <http://www.aallnet.org/chapter/llops/committees/internet/google.htm>.

Happy Googling!