

# **Google Guide**

# Making Searching Even Easier

Nancy Blachman nancy@googleguide.com

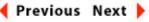
The fact that Nancy has been teaching Internet novices is apparent. She takes nothing for granted, and even includes tips on how to navigate a Web page. More savvy users may skip those sections, however, and focus on the practical examples and exercises.

She is very thorough, and includes introductions to advanced Boolean searching, as well as many of the additional features available at Google, including everything from Froogle shopping search to the Google toolbar.

--Pandia Search Engine News

Google Guide offers help with searching Google, for both the novice and the advanced user. ... The most useful search guides like this one always seems to come from the fringes, not from the company itself.

--LibrarianInBlack.net, resources and discussions for the "tech-librarians-y-default" among us



## **Power Googling: Getting What You Want from Google**

In this presentation, you can learn

- How to select terms and search (more) effectively
- How Google interprets your query
- What's included with your results
- How Google works
- How to invoke search operators (aka advanced operators), even undocumented ones
- When Google added services, features, and options (Google's Feature History)

# **Selecting Search Terms**

The search terms you enter and the order in which you enter them affect both the order and

pages that appear in your search results. In the examples below, click on the similar ways of specifying various searches and note how the results differ.

For simplicity sake, this tutorial uses square brackets to denote Google's search box. For example, to search for a cheap hotel in Mykonos, I'll put the words "cheap," "hotel," and "Mykonos" in square brackets, [ cheap hotel Mykonos ], to indicate you should type those three words in Google's search box. You should not type the brackets, although Google will ignore them if you do type them.

Furthermore, in the examples that follow, each set of search terms is linked to the results of a Google search on those terms. So clicking on [ <a href="mailto:cheap Mykonos hotel">cheap Mykonos hotel</a>] returns the Google results page for a search on those three words.

#### Use words likely to appear on the pages you want.

```
[ salary negotiation tips ]
[ sciatica ]
[ window treatments ]
```

Avoid using a question as a query. For example, the query, [ where do I apply for a passport in New Zealand ], instructs Google to find pages containing all the terms. Such a query won't necessarily find pages answering your question. A better query might be [ passport apply New Zealand ].

```
USE [ passport apply New Zealand ]
NOT [ where do I apply for a passport in New Zealand ]
```

When Google detects very common words such as *where*, *do*, *I*, *for*, and *a*, known as *stop words*, it ignores them so Google may return relevant results. If you're seeking pages that include a stop word, e.g., "how the west was won," learn how to force Google to search for a <u>complete phrase</u> or a <u>specific word</u> in the section <u>Crafting Your Query</u>.

Avoid using words that you might associate with your topic, but you wouldn't expect to find on the designated page(s). For example, queries that include "articles about," "discussion of," "documentation on," and "pages about" are likely to return fewer results since information on the web is rarely labeled with such terms.

```
USE [ lasik eye surgery ]
NOT [ documentation on lasik eye surgery ]

USE [ jobs product marketing Sunnyvale ]
NOT [ listings of product marketing jobs in Sunnyvale ]
```

Suppose you want to know how old someone is, such as Nelson Mandela (the former President of South Africa). Pages with "birthday" or "age" might be more than a year old. Searching for pages that include "Nelson Mandela" and "born" are likely to include either "Nelson Mandela"

born" or "Nelson Mandela was born" followed by his birth date. You can figure out his age from knowing when he was born.

USE [ Nelson Mandela born ]
NOT [ Nelson Mandela birthday ] nor [ Nelson Mandela age ]

Not sure what word or phrase is likely to appear on pages you want? Consider running a word or phrase popularity contest with <u>Google Smackdown</u>, which you can find at www.onfocus.com/ googlesmack/down.asp. This third-party application reports which of two terms or phrases Google estimates to be more prevalent on the web (actually on more web pages that Google has included in its index).



Although not as popular according to Google Smackdown, in Google Guide I use "screen shot" because it's in my online dictionary and "screenshot" isn't.

*Note:* The section <u>How Google Works</u> describes how Google finds web pages and constructs an index.

Be specific: Use more query terms to narrow your results.

It's better to use a more precise, less ambiguous term than a common one to "flesh out the topic by including facets that interest you," notes Ned Fielden in his book *Internet Research*, *Second Edition* (McFarland & Company, 2001).

Does your query have enough specific information for Google to determine unambiguously what you're seeking? If your query is too vague, it's unlikely to return relevant results. Consider, for example, the query [ java ]. What do you suppose Google includes in the first page of results? An island in Indonesia? A beverage consisting of an infusion of ground coffee beans? A computer network-oriented platform-independent programming language developed by Sun Microsystems?

```
USE [ <u>Java Indonesia</u> ], [ <u>java coffee</u> ], or [ <u>java programming language</u> ] NOT [ <u>java</u> ]
```

How can you come up with more specific search terms? What do you know about the topic? Consider answers to the questions, who?, what?, where?, when?, why?, and how?

When you search for [ Tom Watson ], on the first page of results you may get references to a member of Parliament, the golfer, the IBM executive, and a Populist Party candidate for President in 1900 and 1904. If you're searching for something that could return many different types of results, you should add a term that distinguishes among them. This way you'll get only results about the specific Tom Watson you're interested in.

```
USE [ Tom Watson MP ], [ Tom Watson golf ], or [ Tom Watson IBM ]

NOT [ Tom Watson ]

USE [ baby development ] or [ baby milestones ]

NOT [ babies ]

USE [ Betty Ford Center drug addiction ]

NOT [ Ford Center ]
```

Note: Google limits queries to 32 words.

#### Be brief.

For best results, use a few precise words. For example, a program on quitting smoking is more likely to include the terms "quit smoking program" than the words "program on quitting tobacco cigarette smoking addiction."

```
USE [ <u>quit smoking program</u> ]
NOT [ <u>program on quitting tobacco cigarette smoking addiction</u> ]
```

#### You don't have to correct your spelling.

There's a good chance that Google will recognize your mistakes and suggest an alternative more

common spelling, usually faster than you can look up the term in an online dictionary.

When you enter: [ Anna Kornikova tennis ]

Google responds: Did you mean: Anna Kournikova tennis

*Note:* Before clicking on Google's suggested spelling, consider whether it's what you want. Spelling checkers, like people, make mistakes.

For more information on Google's spelling correction system, see the section <a href="Spelling Corrections">Spelling Corrections</a>.

*Note:* Even if you use the search tips described in Google Guide, you won't be able to access authoritative information that's available offline, e.g., old reference books, or is stored in specialized databases. For such information is not currently searchable with Google.

Next we'll look at how Google interprets your query.

For more information on the basics of Google search, visit www.google.com/help/basics.html.

#### **Exercises**

These problems give you practice in selecting search terms. For hints and answers to selected problems, see the <u>Solutions</u> page in the Appendix.

- 1. Find a page with "Google doodle."
- 2. Find the Dilbert cartoon that Scott Adams developed by using Google's logo.
- 3. What's Google's history?
- 4. Find contact information for your representative(s), e.g., senator, congresswoman (or congressman), or member of Parliament.
- 5. How long did it take the first person to cross the United States by car and in what year was it first done?
- 6. In the summer of 1997, an email message was widely circulated featuring the text of a "commencement speech" purportedly given by Kurt Vonnegut at MIT. The imaginary speech began "Wear sunscreen." What's the story behind this email hoax? What did this funny well-written fantasy "commencement speech" say?
- 7. Learn about the recommended tours of the Hearst Castle.
- 8. Find a recipe for lamb with mint sauce.

#### **Interpreting Your Query**

[an error occurred while processing this directive]

#### What Appears on the Results Page

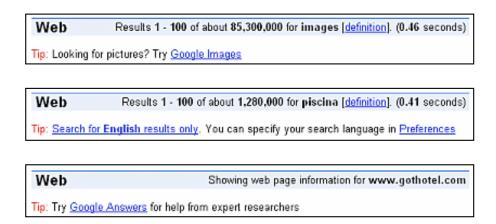
The results page is filled with information and links, most of which relate to your query.



- Google Logo: Click on the Google logo to go to Google's home page.
- Statistics Bar: Describes your search, includes the number of results on the current results page and an estimate of the total number of results, as well as the time your search took. For the sake of efficiency, Google estimates the number of results; it would take considerably longer to compute the exact number. This estimate is unreliable.

Every underlined term in the statistics bar is linked to its <u>dictionary definition</u>. Queries that are linked to just one definition are followed by a definition link.

• **Tips:** Sometimes Google displays a tip in a box just below the statistics bar.



Search Results: Ordered by relevance to your query, with the result that Google considers
the most relevant listed first. Consequently you are likely to find what you're seeking quickly
by looking at the results in the order in which they appear. Google assesses relevance by
considering over a hundred factors, including how many other pages link to the page, the
positions of the search terms within the page, and the proximity of the search terms to one
another.

Below are descriptions of some search-result components. These components appear in fonts of different colors on the result page to make it easier to distinguish them from one another.

- Page Title: (blue) The web page's title, if the page has one, or its URL if the page has no title or if Google has not indexed all of the page's content. Click on the page title (e.g., The History of the Brassiere - Mary Phelps Jacob) to display the corresponding page.
- Snippets: (black) Each search result usually includes one or more short excerpts of the text that matches your query with your search terms in **boldface** type. Each distinct excerpt or snippet is separated by an ellipsis (...). These snippets, which appear in a black font, may provide you with
  - The information you are seeking
  - What you might find on the linked page
  - Ideas of terms to use in your subsequent searches

When Google hasn't crawled a page, it doesn't include a snippet. A page might not be crawled because its publisher requested no crawling, or because the page was written in such a way that it was too difficult to crawl.

- URL of Result: (green) Web address of the search result. In the screen shot, the URL of the first result is inventors.about.com/library/weekly/aa042597.htm.
- Size: (green) The size of the text portion of the web page. It is omitted for sites not yet indexed. In the screen shot, "5k" means that the text portion of the web page is 5 kilobytes. One kilobyte is 1,024 (2<sup>10</sup>) bytes. One byte typically holds one character. In general, the average size of a word is six characters. So each 1k of text is about 170 words. A page containing 5K characters thus is about 850 words long.

Large web pages are far less likely to be relevant to your query than smaller pages. For the sake of efficiency, Google searches only the first 101 kilobytes (approximately 17,000 words) of a web page and the first 120 kilobytes of a pdf file. Assuming 15 words per line and 50 lines per page, Google searches the first 22 pages of a web page and the first 26 pages of a pdf file. If a page is larger, Google will list the page as being 101 kilobytes or 120 kilobytes for a pdf file. This means that Google's results won't reference any part of a web page beyond its first 101 kilobytes or any part of a pdf file beyond the first 120 kilobytes.

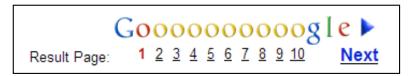
Date: (green) Sometimes the date Google <u>crawled</u> a page appears just after the size
of the page. The date tells you the freshness of Google's copy of the page. Dates
are included for pages that have recently had a <u>fresh crawl</u>.

Indented Result: When Google finds multiple results from the same website, it lists the most relevant result first with the second most relevant page from that same site indented below it. In the screen shot, the indented result and the one above it are both from the site inventors.about.com.

Limiting the number of results from a given site to two ensures that pages from one site will not dominate your search results and that Google provides pages from a variety of sites.

More Results: When there are more than two results from the same site, access the remaining results from the "More results from..." link.

When Google returns more than one page of results, you can view subsequent pages by clicking either a page number or one of the "o"s in the whimsical "Gooooogle" that appears below the last search result on the page.



If you find yourself scrolling through pages of results, consider <u>increasing the number</u> <u>of results Google displays on each results page</u> by changing your global preferences (see the section Changing Your Global Preferences).

In practice, however, if pages of interest to you aren't within the first 10 results, consider refining your query instead of sifting through pages of irrelevant results. To simplify such refinements, Google includes a search box at the bottom of the page you can use to enter your refined query.

- Sponsored Links: Your results may include some clearly identified sponsored links (advertisements) relevant to your search. If any of your search terms appear in the ads, Google displays them in **boldface** type.
- Spelling Corrections, Dictionary Definition, Cached, Similar Pages, News, Product Information, Translation, Book results: Your results may include these links, which are described on the next few pages.

Here's another screen shot of the results page in case the one at the top of this page scrolled off your screen.



For more on what's included on Google's results page, visit <a href="https://www.google.com/help/interpret.html">www.google.com/help/interpret.html</a>.

#### **How Google Works**

If you aren't interested in learning how Google creates the index and the database of documents that it accesses when processing a query, skip this description. I adapted the following overview from Chris Sherman and Gary Price's wonderful description of How Search Engines Work in Chapter 2 of <a href="The Invisible Web">The Invisible Web</a> (CyberAge Books, 2001).

Google runs on a distributed network of thousands of low-cost computers and can therefore carry out fast parallel processing. Parallel processing is a method of computation in which many calculations can be performed simultaneously, significantly speeding up data processing. Google has three distinct parts:

- Googlebot, a web crawler that finds and fetches web pages.
- The indexer that sorts every word on every page and stores the resulting index of words in a huge database.
- The query processor, which compares your search query to the index and recommends the documents that it considers most relevant.

Let's take a closer look at each part.

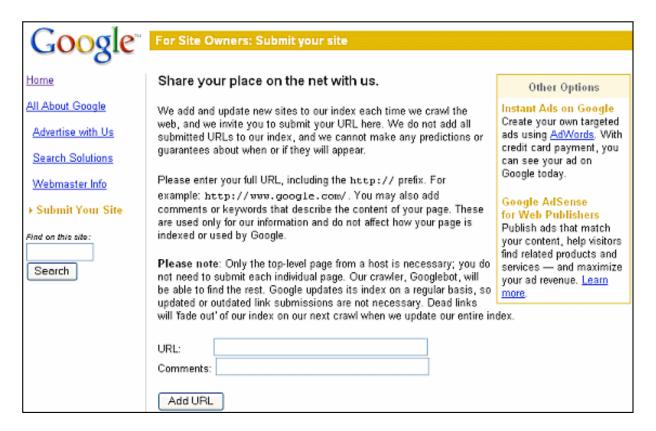
#### Googlebot, Google's web Crawler

Googlebot is Google's web crawling robot, which finds and retrieves pages on the web and hands them off to the Google indexer. It's easy to imagine Googlebot as a little spider scurrying across the strands of cyberspace, but in reality Googlebot doesn't traverse the web at all. It functions much like your web browser, by sending a request to a web server for a web page, downloading the entire page, then handing it off to Google's indexer.

Googlebot consists of many computers requesting and fetching pages much more quickly than

you can with your web browser. In fact, Googlebot can request thousands of different pages simultaneously. To avoid overwhelming web servers, or crowding out requests from human users, Googlebot deliberately makes requests of each individual web server more slowly than it's capable of doing.

Googlebot finds pages in two ways: through an add URL form, <a href="www.google.com/addurl.html">www.google.com/addurl.html</a>, and through finding links by crawling the web.



Unfortunately, spammers figured out how to create automated bots that bombarded the add URL form with millions of URLs pointing to commercial propaganda. Google rejects those URLs submitted through its Add URL form that it suspects are trying to deceive users by employing tactics such as including hidden text or links on a page, stuffing a page with irrelevant words, cloaking (aka bait and switch), using sneaky redirects, creating doorways, domains, or subdomains with substantially similar content, sending automated queries to Google, and linking to bad neighbors. So now the Add URL form also has a test: it displays some squiggly letters designed to fool automated "letter-guessers"; it asks you to enter the letters you see — something like an eye-chart test to stop spambots.

When Googlebot fetches a page, it culls all the links appearing on the page and adds them to a queue for subsequent crawling. Googlebot tends to encounter little spam because most web authors link only to what they believe are high-quality pages. By harvesting links from every page it encounters, Googlebot can quickly build a list of links that can cover broad reaches of the web. This technique, known as deep crawling, also allows Googlebot to probe deep within individual sites. Because of their massive scale, deep crawls can reach almost every page in the web. Because the web is vast, this can take some time, so some pages may be crawled only once a month.

Although its function is simple, Googlebot must be programmed to handle several challenges. First, since Googlebot sends out simultaneous requests for thousands of pages, the queue of "visit soon" URLs must be constantly examined and compared with URLs already in Google's index. Duplicates in the queue must be eliminated to prevent Googlebot from fetching the same page again. Googlebot must determine how often to revisit a page. On the one hand, it's a waste of resources to re-index an unchanged page. On the other hand, Google wants to re-index changed pages to deliver up-to-date results.

To keep the index current, Google continuously recrawls popular frequently changing web pages at a rate roughly proportional to how often the pages change. Such crawls keep an index current and are known as *fresh crawls*. Newspaper pages are downloaded daily, pages with stock quotes are downloaded much more frequently. Of course, fresh crawls return fewer pages than the deep crawl. The combination of the two types of crawls allows Google to both make efficient use of its resources and keep its index reasonably current.

#### Google's Indexer

Googlebot gives the indexer the full text of the pages it finds. These pages are stored in Google's index database. This index is sorted alphabetically by search term, with each index entry storing a list of documents in which the term appears and the location within the text where it occurs. This data structure allows rapid access to documents that contain user query terms.

To improve search performance, Google ignores (doesn't index) common words called *stop* words (such as *the*, *is*, *on*, *or*, *of*, *how*, *why*, as well as certain single digits and single letters). Stop words are so common that they do little to narrow a search, and therefore they can safely be discarded. The indexer also ignores some punctuation and multiple spaces, as well as converting all letters to lowercase, to improve Google's performance.

#### **Google's Query Processor**

The query processor has several parts, including the user interface (search box), the "engine" that evaluates queries and matches them to relevant documents, and the results formatter.

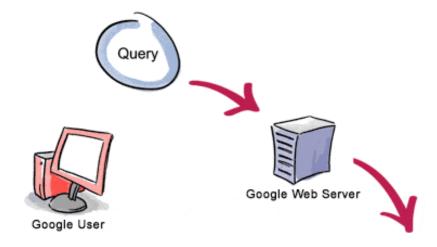
Google considers over a hundred factors in determining which documents are most relevant to a query, including the popularity of the page, the position and size of the search terms within the page, and the proximity of the search terms to one another on the page. <a href="PageRank">PageRank</a> is Google's system for ranking web pages. (View a websites's PageRank from Google's ToolBar and from the Google Directory.)

Google also applies machine-learning techniques to improve its performance automatically by learning relationships and associations within the stored data. For example, the <a href="mailto:spelling-correcting-system">spelling-correcting-system</a> uses such techniques to figure out likely alternative spellings. Google closely guards the formulas it uses to calculate relevance; they're tweaked to improve quality and performance, and to outwit the latest devious techniques used by spammers.

Indexing the full text of the web allows Google to go beyond simply matching single search terms. Google gives more priority to pages that have search terms near each other and in the same

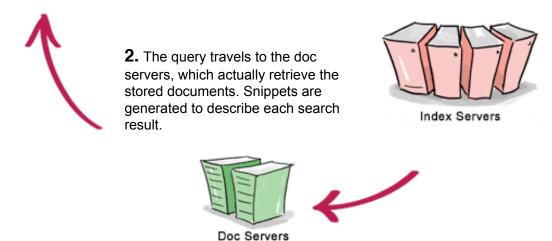
order as the query. Google can also match multi-word phrases and sentences. Since Google indexes HTML code in addition to the text on the page, users can restrict searches on the basis of where query words appear, e.g., in the title, in the URL, in the body, and in links to the page, options offered by the <u>Advanced-Search page</u> and <u>search operators</u>.

Let's see how Google processes a query.



**3.** The search results are returned to the user in a fraction of a second.

**1.** The web server sends the query to the index servers. The content inside the index servers is similar to the index in the back of a book--it tells which pages contain the words that match any particular query term.



Copyright © 2003 Google Inc. Used with permission.

### **Feature History**

Google is constantly enhancing its search engine. The following table lists when Google and Google Guide added services and features and links to where they are described in Google Guide or on the web.

September 2005	Google Blog Search	
August 2005	Google Desktop 2	Google Talk     Useful links from within a site
July 2005	Toolbar for Firefox	
June 2005	Currency Conversion     Google Earth	Google Sitemaps
May 2005	<ul><li>Enterprise Desktop Search Tool</li><li>Google Web Accelerator</li></ul>	Personalize Your Homepage
April 2005	<ul><li> Google Q&amp;A</li><li> Multilingual Google Definitions</li><li> My Search History</li></ul>	<ul> <li>Satellite View from Google Maps</li> <li>Google Mobile - Local Search</li> </ul>
March 2005	<ul><li>Google Weather</li><li>Live stock quotes</li><li>Results Prefetching</li></ul>	<ul><li>Google News: Personalized News</li><li>Google Ride Finder</li><li>Google Suggest in Japanese</li></ul>
February 2005	• Google Maps	Google Movies
January 2005	Raised search limit to 32 words     Picasa 2	• Google Video
December 2004	Google Library     Froogle Product Reviews	Google Suggest
November 2004	Google Scholar     Froogle Wish Lists	Google Help: Cheat Sheet
October 2004	Google Desktop Search	• Google SMS
June 2004	Site-Flavored Google Search Box	

May 2004	<ul> <li>Google Groups 2</li> <li>Added mailing list support to Google Groups</li> </ul>	
April 2004	• GMail	
March 2004	<ul> <li>Google Local</li> <li>User Interface (UI) redesign</li> <li>Personalized Web Search</li> <li>Web Alerts</li> </ul>	<ul> <li>New Froogle home page</li> <li>Images in Google News search results</li> <li>Number range (numrange) command</li> </ul>
February 2004	Danish Google Guide	
January 2004	Search by Number	Orkut (Social Networking Service)
December 2003	Travel Conditions  Google Book Search	<ul><li> Product Search Shortcut</li><li> Word Variation (Automatic Stemming)</li></ul>
November 2003	• <u>Deskbar</u>	
October 2003	Definitions (Google Glossary)	
September 2003	Search by Location	
August 2003	<ul><li><u>Calculator</u></li><li>Synonym Operator</li></ul>	• News Alerts
June 2003	Toolbar 2.0 with a pop-up blocker	
December 2002	• Froogle	• Google Viewer
September 2002	• Google News	
May 2002	Google Answers Searchable     Google Labs	Google Glossary     Google Sets
April 2002	Google Answers	
Jan 2002	* Wildcard (currently not working)	

December 2001	+ Operator • Catalogs	• <u>Diacritics searching</u> (terms with accents)
November 2001	Search stop words in phrases	• File type conversion
October 2001	Home page tabs	• Language Tools
June 2001	Google Image Search	
May 2001	Spelling Corrections     Google Groups	• <u>Translation</u>
March 2001	Phonebook	
November 2000	• <u>Toolbar</u>	
October 2000	Stock Quotes  • AdWords	OR Operators
July 2000	<ul><li>Date Restricted Search</li><li>Dictionary Link</li><li>Advanced Search</li></ul>	Preferences     Phrase Search     Operator
June 2000	News Headlines	Street Maps
May 2000	SafeSearch Filtering	
April 2000	Google Directory	
March 2000	Browser Buttons	
February 2000	Microsoft-Specific Search Engine	
January 2000	Make Google Your Default Engine     Apple-Specific Search Engine	University Search Feature
September 1999	Similar Pages aka GoogleScout	
September 1998	• <u>Cached Pages</u>	Google Web Search

Resources helpful in putting together this table include <u>Google Press Releases</u>, the <u>archive of Google-Friends Newsletters</u>, <u>Google: Search Engine Showdown News Archive</u>, and <u>Google News</u>.

This page was last modified on Monday January 09, 2006.

[Home] [Intro] [Contents] [Print] [Favorites] [Query Input] [Understanding Results] [Special Tools] [Developing a Website] [Appendix]







Google Search

Let me know your thoughts on Google Guide by emailing nancy(at)googleguide.com (replace (at) by @).

For permission to copy and create derivative works, visit the <u>Creative Commons License</u> page. Google Guide is neither affiliated with nor endorsed by Google.