Upper Arlington High School Journalism
Hemmerly

Investigative Reporting & Technology

Searching the WWW
Computer-Assisted Reporting
Using Statistics

What Makes a Search Engine Good?

All search engines consist of three parts: (1) a database of web documents, (2) a search engine operating on that database, and (3) a series of programs that determine how search results are displayed. Because the search engine business is competitive, most search engines also offer additional features that are convenient or fun. The table below shows what can vary within each of the three basic parts in search engines.

Parts of Search Engines	Variables, and their implications for your searches

1. Database of web documents	• **Size of database:**
	  ○ How many documents does the search engine claim it has?
	  ○ How much of the total web are you able to search?
	• **Freshness** ("up-to-dateness"):
	  ○ Search engine databases consist of copies of web pages and other documents that were made when their crawlers or spiders last visited each site. How often is the database refreshed to find new pages?
	  ○ How often do their crawlers update the copies of the web pages you are searching?
	• **Completeness of text:**
	  ○ Is the database really "full" text, or only parts of the pages?
	  ○ Is every word indexed?
	• **Types of documents offered:**
	  ○ All search engines offer web pages.
	  ○ Do they also have extensive PDF, Word, Excel, PowerPoint, and other formats like WordPerfect?
	  ○ Are they full-text searchable?
	• **Speed and consistency:**
	  ○ How fast is it?
	  ○ How consistent is it? Do you get different results at different times?

2. The search engine's capabilities
All search engines let you enter some keywords and search on them. What happens inside?
Can you limit in ways that will increase your chances of finding what you are looking for?
	• **Basic Search options and limitations:**
	  ○ Automatic default of AND assumed between words?
	  ○ Accepts " " to create phrases?
	  ○ Is there an easy way to allow for synonyms and equivalent terms (OR searching)? Can you OR phrases or just single words?
	• **Advanced Search options and limitations:**
	  ○ Can you require your search terms in specific fields, such as the document title? Can you require some words in certain fields and others anywhere?
	  ○ Can you restrict to documents only from a certain domain (org, edu, gov, etc.)? Limit to more than one or only one?
	  ○ Can you limit by type of document (pdf or excel, etc.)? More than one?
	  ○ Can you limit by language?
	  ○ How reliably and easily can you limit to date last updated?
	• **General limitations and features:**
	  ○ What do you have to do make it search on common or stop words?
	  ○ Maximum limit on search terms or on search complexity?
	  ○ Ability to search within previous results?
	  ○ Can you count on consistent results from search to search and from day to day?
	  ○ Can you customize the search or display?
	  ○ Is there a "family" filter? Does it work well? Is it easy to turn on or off?

3. Results display
All search engines return a list of results it "thinks" are what you are looking for. How well does it "think like you expect it think"?
	• **Ranking:**
	  ○ Are they ranked by popularity or relevancy or both?
	  ○ Do pages with your words juxtaposed (like a phrase) rank highest?
	  ○ Do you get pages with only some of your words, perhaps in addition to pages with them all?
	• **Display:**
	  ○ Are your keywords highlighted in context, showing excerpts from the web pages which caused the match?
	  ○ Some other excerpt from the page?
	• **Collapse pages from the same site:**
	  ○ If it shows only one or a few pages from a site, does it show the one(s) with your terms?
	  ○ How easy is it to see all from the site?
	  ○ Can this be changed and saved as your preferred search method?

4. Other features	Search engine designers try to come up with all kinds of features and services that they hope will allure you to their services.

Recommended Search Engines: Table of Features

Google has one of the largest databases of Web pages, including many other types of Web documents (e.g., PDFs, Word or Excel documents, PowerPoints). Despite the presence of many advertisements and considerable clutter from blog sites and newsgroups, Google's popularity ranking often makes pages worth looking at rise near the top of search results. Google alone is not sufficient, however. Less than half the searchable Web is fully searchable in Google.

Overlap studies show that about half of the pages in any search engine database exist only in that database. Getting a second opinion is therefore often worthwhile. For a second opinion, we recommend Teoma, Vivisimo (a meta-search engine that indirectly searches three huge search engine databases) or Yahoo! Search.

Search Engine	Google www.google.com	Yahoo! Search search.yahoo.com	Teoma www.teoma.com
Links to help	Google help pages	Yahoo! help pages	Teoma help pages
Size, type Size varies frequently and widely. See tests and more charts.	HUGE. Over 3 billion. Claims over 4 billion but about 1 billion are not fully indexed (i.e., cannot be full-text word searched). Unindexed pages are retrieved if your search matches their titles or match other pages linking to them. Biggest in tests.	HUGE. Over 3 billion fully indexed, searchable pages.	LARGE. Claims to have 1 billion fully indexed, searchable pages, and 1 billion more partially indexed. Strives to become #1 in size.
Noteworthy features and limitations	Popularity ranking using PageRank™. Limit of 10 words per search, excluding OR. Indexes the first 101KB of a Web page, and 120KB of PDF's.	Shortcuts give quick access to dictionary, synonyms, patents, traffic, stocks, encyclopedia, and more.	Subject-Specific Popularity™ ranking. Suggests terms within results to refine Suggests pages within results with many links.
Phrase searching (term definition)	Yes. Use " ". Searches common "stop words" if in phrases in quotes.	Yes. Use " "	Yes. Use " ". Searches common "stop words" if in phrases in quotes.
Boolean logic (term definition)	Partial. AND assumed between words. Capitalize OR. - excludes. No ( ) or nesting. In Advanced Search, partial Boolean available in boxes.	Accepts AND, OR, NOT or AND NOT, and ( ). *Must be capitalized.*	Partial. AND assumed between words. Capitalize OR. - excludes. No ( ) or nesting.
+Requires/ -Excludes (term definition)	- excludes + will allow you to retrieve "stop words" (e.g., +in)	- excludes + will allow you to search common words: "+in truth"	- excludes + will allow you to retrieve "stop words" (e.g., +in)
Sub-Searching (term definition)	Sort of . At bottom of results page, click "Search within results" and enter more terms. Adds terms.	Add terms.	Sort of . Add terms. REFINE pastes suggested sub-topics within results.
Results Ranking (term definition)	Based on page popularity measured in links to it from other pages: high rank if a lot of other pages link to it. Fuzzy AND also invoked. Matching and ranking based on "cached" version of pages that may not be the most recent version.	Automatic Fuzzy AND.	Based on Subject-Specific Popularity™, links to a page by related pages. More info.
Field limiting (term definition)	link: site: allintitle: intitle: allinurl: inurl: Advanced Search boxes for most of these. Offers Uncle Sam for US federal pages and other special searches.	link: site: intitle: inurl: url: hostname: (Explanation of these distinctions.)	intitle: inurl: site: geoloc: Explanations, limitations.
Truncation (term definition)	No. Search variant endings and synonyms separately, separating with OR (capitalized): *airline OR airlines*	No. Search with OR as in Google.	No. Search with OR as in Google.
Case sensitivity (term definition)	No.	No.	No.
Language	Yes. Major Romanized and non-Romanized languages in Advanced Search.	Yes. Major Romanized and non-Romanized languages.	Yes. Major Romanized languages. Use lang:
Limit by age of documents	In Advanced Search.	In Advanced Search.	In Advanced Search.
Translation	Yes, in Translate this page link following some pages. To English from major European languages.	No.	No.

The Boolean Search

Don't let the name scare you, this isn't that difficult to understand, it just requires getting used to.

There are many ways of searching for documents, papers, books, etc. on a given topic, but most of them require the information to be better organized and more standardized than documents on the web are. As a result, users like yourself are stuck with doing a lot of free-text searching, meaning, looking for documents that contain words you think will be in the document you are seeking.

Most search engines give you the option of entering a boolean search, that is, one that uses AND and OR and looks something like this: "White house" and (schedule or calendar) and (events or tours). It is really very difficult to have successful free-text searches without understanding how to build search strings like this. You really do not have enough power to narrow your search to a reasonable number of potentially useful documents without it.

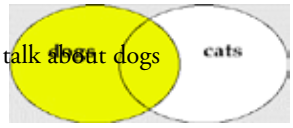
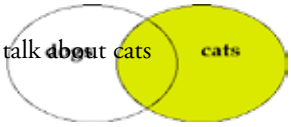
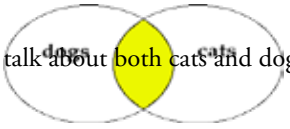
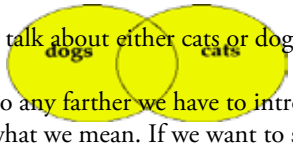
Boolean basics

Venn Diagrams
Boolean concepts are often explained with Venn diagrams (the circle plots below) the Venn diagrams in this tutorial are meant to represent the information space of the web. A circle with a word in it shows the subset of web documents that contain that concept. When there are two concepts shown in a single diagram, the overlap of the circles represents documents that contain both concepts. For example:
The diagram above shows the information space of documents that talk about dogs and cats. Some documents, the blue region, talk about dogs and not cats, some documents, the yellow region, talk about cats and not dogs, and some talk about both (the green region).

Basic AND and OR
AND and OR in the context of boolean sometimes seem to mean the opposite that they do in language, so try not to think of boolean strings as English sentences.

AND means "I want only documents that contain both words."
OR means "I want documents that contain either word. I don't care which word."

So going back to the cat and dog diagram. If I want the yellow areas in each of the following diag string shown next to it.

documents that talk about dogs 	boolean search: dogs
documents that talk about cats 	boolean search: cats
documents that talk about both cats and dogs 	boolean search: cats AND dogs
documents that talk about either cats or dogs 	boolean search: cats OR dogs

Parenthesis ( )
Before we can go any farther we have to introduce a concept you may have seen in algebra: using parenthesis to make sure the computer knows exactly what we mean. If we want to search for more than two concepts, or use AND and OR in a search, we have to tell the com-

puter what part of the search to execute first. Items contained within parentesis ( ) are always intrepreted or executed first.

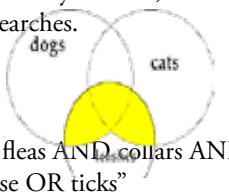
For Example: if I look for articles about using leashes with dogs or using leashes with cats with the search "cats OR dogs AND leashes" I may not retrieve the documents I intend. The computer does not read from left to right the way humans do. In fact it has a completely different way of looking at this search; most search engines interpret the ANDs first followed by the ORs. So what I would really get out of this search is documents that talk about dogs with leashes along with any document about cats! Why? The diagrams below will demonstrate:

Our information space looks like this:

If we interpret the search "cats OR dogs AND leas mputer does, we would first AND the circles for dogs and leashes, the yellow area. Then we would OR that resulting are ats resulting in documents in both the yellow and orange areas.

To solve this problem we tell the computer that the OR is to be interpreted first. So instead we use the search string "(dogs OR cats) AND leashes". This string would result in the yellow information s low being retrieved.

Even when you think the computer is going to do what you want, it is always safer to use parenthesis if there is even a chance of confusion. The parenthesis will also help you read your own searches.



There are times when parenthesis are not needed:

* Only using AND: "dogs AND fur AND fleas AND collars AND rugs"
* Only using OR: "fleas OR gnats OR tsetse OR ticks"

Intermediate AND and OR
We are going to get into some more difficult searches now using some more complex information spaces. Hopefully by the end of this section you will be able to construct searches for most every situation you ever encounter in your Web searches. Some of these may seem a little tough at first, but take it slow and study the diagrams. Sometimes it may help to imagine what kind of documents could fall into each part of the diagrams.

NEAR
With AND the terms can appear anywhere in the document. Within a long document a lot of different words will create combinations that are not really discussed in the document. For example a certain document may have "white barn" in the first paragraph and "red wagons" twenty paragraphs later, if I do a search for "red AND barns" I will get this document even though it has nothing to do with red barns or barns that are painted red.

So instead of the above search, I really want to use "red NEAR barns". This means I will get documents with sentences like "Barry headed down the path to the red barn" or "we took some red clothes out of the barn". The tolerance of NEAR varies by search engine with a range 9 to 15 words being typical. Note that sometimes this feature is called WITHIN, but the three below all use the syntax NEAR.

* AltaVista: 10 words
* WebCrawler: user controlled ("red NEAR/15 barn")

Quoted Strings: or how to really narrow your search
[discuss these with known item searches in four searches doc.] In the above red barn example I had the option of replacing red NEAR barn with "red barn" in quotes to force the search engine to treat it as a literal string. This really would have reduced the number of documents I retrieved. It may have reduced it too much, there are many ways of writing about red barns.

NOT, use it sparingly
What does NOT do? NOT tells the search engine to throw out any documents that contain that word. This command is usually too power-

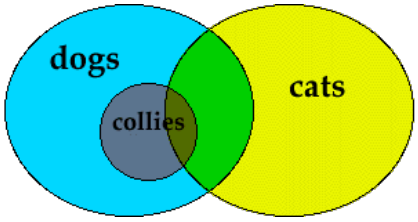
ful to use.

For Example: say while I was doing my leash research I kept getting a whole bunch of documents about people who went for walks with their dogs and their llamas. Say also that all of these authors were obsessed with their new llama leashes and never seemed to get around to talking about dog leashes. I may be tempted to change my search to ignore any document that contains the word "llama". But I may be eliminating the very documents I really want to get. What if the foremost expert on leashes always dedicates her papers to her pet llama, or her name is Mildred P. Llama. The NOT directive is completely non-discriminatory; it only takes one single instance of a word to eliminate a document from your retrieved set.

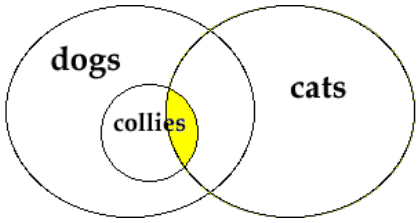
Be aware of redundant terms
Sometimes you may find yourself creating searches with redundant terms. They aren't really harmful, but it is important that you understand what is happening, so if the search isn't bringing back the documents you expect, you can edit it properly.

Say I am searching in the information space shown in this diagram:

Say I want documents about collies and cats growing up together, the yellow area shown here:



I may be tempted to use the search: dogs AND cats AND collies but the word "dogs" is redundant since all of the documents about collies are also about dogs. The search cats AND collies gets me the same yellow area.



A good general rule (for your first few search attempts at least) is to stick with a simple search and resort to a more complicated one only if you are not finding what you want.

Source: http://florin.syr.edu/webarch/searchpro/boolean_tutorial.html